ATPBR

DOCUMENTAÇÃO DA PLATAFORMA

Atlas Tumoral da População Brasileira

Uma iniciativa dedicada à compreensão da heterogeneidade biológica e genômica do câncer na população brasileira.



atpbr.com.br

Instituto Nacional de Câncer (INCA) · Versão 0.1 · novembro de 2026

Sobre este livreto

Este livreto reúne a documentação da plataforma do Atlas Tumoral da População Brasileira (ATPBR): o que é o projeto, como os dados são governados, como submeter e acessar dados, como as informações se organizam e como os resultados são processados e explorados.

Os capítulos seguem o ciclo de vida dos dados na plataforma e estão agrupados em quatro partes. Cada capítulo pode ser lido de forma independente: as referências cruzadas indicam o número do capítulo em que cada assunto é detalhado.

O conteúdo corresponde à documentação online do ATPBR na data desta edição. As regras de governança e de acesso seguem as políticas institucionais do INCA; em caso de divergência, prevalecem os documentos institucionais vigentes.

TÍTULO	Atlas Tumoral da População Brasileira (ATPBR): documentação da plataforma
VERSÃO	Versão 0.1 · novembro de 2026
INSTITUIÇÃO	Instituto Nacional de Câncer (INCA), Ministério da Saúde
SITE	www.atpbr.com.br
CONTATO	atpbr@inca.gov.br
COMO CITAR O ATPBR	Boroni M, Costa Conteville L, Ribeiro Pinto LF, P B Viola J. *The Brazilian Population Tumor Atlas Project: addressing representation in cancer genomics in Brazil*. Lancet Reg Health Am. 2026;56:101397. doi:10.1016/j.lana.2026.101397



PALAVRA DA COORDENAÇÃO

Dados brasileiros para a oncologia brasileira

A medicina de precisão já mudou a forma como diagnosticamos e tratamos muitos tipos de câncer. Ela depende, porém, de dados, e os dados que hoje a sustentam pouco se parecem com a população que o Sistema Único de Saúde atende.

A maior parte do conhecimento em genômica do câncer foi produzida com pacientes do Hemisfério Norte. No The Cancer Genome Atlas (TCGA), um dos maiores repositórios do mundo, 77% das amostras vieram de pacientes brancos e apenas 3% de pacientes hispânicos ou latinos. A população brasileira, uma das mais diversas do planeta, com contribuições indígenas, africanas e europeias, segue quase ausente dessas bases.

Essa ausência tem consequências concretas. Um estudo com 2.723 genomas brasileiros encontrou mais de 8,7 milhões de variantes genéticas nunca descritas. Critérios de testagem genética usados no país deixariam de identificar entre 17% e 25% das pessoas com variantes associadas ao câncer hereditário. Variantes que definem risco, prognóstico e resposta ao tratamento não se transferem automaticamente de uma população para outra, e decisões clínicas apoiadas em dados pouco representativos tendem a ser menos eficazes para os brasileiros.

O Atlas Tumoral da População Brasileira (ATPBR) nasce para enfrentar essa lacuna. Liderado pelo INCA e pelo Consórcio Genoma Câncer Brasileiro, com financiamento do Ministério da Saúde em parceria com o CNPq, o projeto reúne dados clínicos e moleculares de pacientes de diferentes regiões do país em uma infraestrutura nacional, segura e governada, que permite integrá-los, compará-los e reutilizá-los.

Ampliar a presença de brasileiros nos dados de câncer não é apenas uma questão de equidade científica. É condição para descobrir biomarcadores relevantes para a nossa população, desenvolver estratégias de diagnóstico e tratamento que funcionem para ela e levar a medicina de precisão a quem é atendido pelo SUS.

Convido pesquisadores, profissionais de saúde e instituições parceiras a conhecer a plataforma descrita neste livreto e a contribuir com ela. Cada caso submetido com qualidade, consentimento e metadados completos aproxima a oncologia brasileira de um retrato fiel dos seus pacientes.

João P.B. Viola
Coordenador de Pesquisa e Inovação
Instituto Nacional de Câncer (INCA)

Sumário

PARTE 1 **Conhecer o ATPBR** 5

01 Sobre o ATPBR 6

02 Comece aqui 10

03 Arquitetura da plataforma 12

04 Governança de dados 14

PARTE 2 **Enviar e compartilhar dados** 25

05 Submissão de dados 26

06 Acesso aos dados 30

PARTE 3 **Organização dos dados** 35

07 Modelo de dados 36

08 Dicionário de dados 38

09 Sistema de identificação e barcode 43

10 Tipos de dados 54

PARTE 4 **Análise e uso dos dados** 56

11 Plataforma analítica 57

12 Navegador de resultados 61

13 Políticas de uso 65

Referência rápida

14 Perguntas frequentes 67

PARTE 1

Conhecer o ATPBR

O projeto, a organização desta documentação, a arquitetura da plataforma e as regras de governança.

01 Sobre o ATPBR

02 Comece aqui

03 Arquitetura da plataforma

04 Governança de dados

CAPÍTULO 01

Sobre o ATPBR

O que é o ATPBR?

O **Atlas Tumoral da População Brasileira (ATPBR)** é uma iniciativa nacional dedicada a mapear as características moleculares do câncer na população brasileira e a compartilhar esses dados de forma responsável. Financiado pelo **Ministério da Saúde** em parceria com o **CNPq** desde dezembro de 2023, e liderado pelo Instituto Nacional de Câncer (INCA) e pelo Consórcio Genoma Câncer Brasileiro, o projeto reúne, organiza e disponibiliza dados clínicos e **multiômicos** (genômica, transcriptômica e epigenômica) de pacientes oncológicos de diferentes regiões do país.

Na prática, o ATPBR funciona como uma **infraestrutura centralizada e segura** que integra os metadados clínicos e os resultados moleculares, permitindo que dados ômicos de tumores produzidos em diferentes estudos, instituições e tecnologias sejam integrados e comparados de forma consistente.

O problema que o ATPBR ajuda a resolver

A maior parte dos dados genômicos que hoje orientam a **medicina de precisão** foi gerada em populações do Hemisfério Norte.

- Em um dos maiores repositórios mundiais, o The Cancer Genome Atlas (TCGA), **77% das amostras analisadas eram de pacientes brancos** e apenas **3% de pacientes hispânicos/latinos** (Spratt et al., 2016).
- **PCAWG/ICGC–TCGA:** 76,2% dos indivíduos apresentavam ancestralidade europeia e apenas 1,0% foi classificado como Admixed American/Amerindian/Latino (Campbell et al., 2020).
- **AACR Project GENIE:** 72% dos pacientes eram classificados como brancos; pacientes hispânicos apresentavam apenas 57% da representação esperada em relação à população oncológica dos EUA (Pugh et al., 2022; Cheung et al., 2023).
- **NCI Patient-Derived Models Repository:** 75% das amostras são de ancestralidade europeia e menos de 1% de ancestralidade Native/Latin American (Molina-Aguilar e Robles-Espinoza, 2023).

Esses dados evidenciam uma lacuna persistente na representação de populações latino-americanas nos principais recursos genômicos utilizados para descoberta de biomarcadores, desenvolvimento de modelos preditivos e implementação da medicina de precisão.

Isso é um problema porque as características genéticas de uma população não se aplicam automaticamente a outra: variantes associadas ao **risco**, ao **prognóstico** e à **resposta a tratamentos** podem ser diferentes. Decisões clínicas baseadas em dados pouco representativos tendem a ser menos eficazes para os brasileiros.

Por que a diversidade brasileira importa?

A população brasileira apresenta **grande diversidade genética**, resultado de uma história demográfica marcada por **contribuições ancestrais indígenas, africanas, europeias e de outras populações**. Essa diversidade representa uma oportunidade científica ainda pouco explorada:

- Um estudo com **2.723 genomas brasileiros** identificou mais de **8,7 milhões de variantes genéticas previamente não documentadas** (Nunes et al., 2025).
- Os critérios de elegibilidade para testagem genética avaliados em um estudo brasileiro deixariam de identificar entre **17% e 25%** dos portadores de **variantes patogênicas ou provavelmente patogênicas** associadas a síndromes hereditárias de predisposição ao câncer (de Oliveira et al., 2022).
- Algumas variantes germinativas patogênicas associadas ao câncer hereditário são **mais frequentes na população brasileira** do que em bases de referência internacionais (Boroni et al., 2026).
- O **Brazilian Genome Map Oncology Subproject (BGMO)** demonstrou a viabilidade de gerar WGS tumor–germinativo de forma multicêntrica no Brasil, com 275 pacientes recrutados nas cinco regiões do país e uma composição populacional distinta das grandes bases internacionais (Ferreira et al., 2026).

Esses dados mostram que o desafio não é apenas **gerar mais dados genômicos brasileiros**, mas garantir que eles possam ser **padronizados, integrados, governados, compartilhados e reutilizados**. O **ATPBR** foi concebido para preencher essa lacuna, criando uma infraestrutura nacional capaz de transformar dados hoje dispersos em um recurso integrado para pesquisa e medicina de precisão.

Nossa missão

Ampliar a representatividade da população brasileira na pesquisa oncológica, disponibilizando recursos que contribuam para a descoberta de biomarcadores, o desenvolvimento de novas estratégias de diagnóstico e tratamento e, em última instância, a oferta de **medicina de precisão no Sistema Único de Saúde (SUS)**.

Para quem o ATPBR foi criado?

A plataforma foi desenvolvida para atender diferentes perfis de usuários, incluindo:

- Pesquisadores interessados em acessar e analisar dados oncológicos.
- Instituições e pesquisadores que desejam compartilhar dados de forma padronizada e segura.
- Bioinformatas e cientistas de dados em busca de dados para suas análises multiômicas.
- Gestores e coordenadores de projetos que necessitam acompanhar o ciclo de vida dos dados.
- Colaboradores envolvidos em iniciativas nacionais de pesquisa em câncer.

O capítulo **Comece aqui** indica o melhor caminho de leitura para cada perfil.

O que o ATPBR oferece?

O ATPBR reúne, em um único ambiente, recursos que acompanham todo o ciclo de vida dos dados. Cada um deles é detalhado em um capítulo desta documentação:

- **Armazenamento seguro:** infraestrutura nacional para armazenamento e gerenciamento de dados clínicos e moleculares, com mecanismos de proteção e controle de acesso (Arquitetura da plataforma).
- **Governança e proteção de dados:** políticas que estabelecem responsabilidades, critérios de acesso e condições de uso, assegurando a proteção dos participantes e a conformidade com as normas éticas e legais (Governança de dados).
- **Submissão padronizada:** mecanismos para o cadastro de projetos e a submissão estruturada de dados provenientes de diferentes instituições e estudos (Submissão de dados).
- **Integração e harmonização:** padronização de dados clínicos, sociodemográficos e multiômicos, permitindo a integração e comparação de informações produzidas por diferentes tecnologias e instituições (Modelo de dados e Dicionário de dados).
- **Processamento e análise reprodutíveis:** utilização de pipelines bioinformáticos padronizados, automatizados e versionados, garantindo consistência, rastreabilidade e reprodutibilidade das análises (Plataforma analítica).
- **Exploração e visualização amigável e interativa:** ferramentas para consultar, explorar e comparar dados moleculares e clínicos, facilitando a identificação de padrões e a geração de novas hipóteses científicas (Navegador de resultados).
- **Soberania nacional e compartilhamento responsável:** manutenção dos dados sob governança brasileira, com acesso público a informações não sensíveis e mecanismos de acesso controlado para dados sensíveis, respeitando as autorizações de uso, os consentimentos e a legislação vigente (Acesso aos dados).

Compromisso com qualidade e confiança

Para garantir dados confiáveis e reutilizáveis, o ATPBR adota os princípios **FAIR** (Encontráveis, Acessíveis, Interoperáveis e Reutilizáveis) e orienta seu desenvolvimento pelos **padrões internacionais da Global Alliance for Genomics and Health (GA4GH),** que promovem a interoperabilidade, a governança e o compartilhamento responsável de dados genômicos e clínicos. A plataforma estabelece três critérios mínimos para a inclusão de amostras: (i) dados de sequenciamento de alta qualidade, (ii) consentimento informado dos participantes/pacientes (ou dispensa aprovada pela instância ética competente), em conformidade com as normas éticas e regulatórias vigentes, e (iii) um conjunto mínimo de metadados clínicos e sociodemográficos.

Em conformidade com a **Lei Geral de Proteção de Dados Pessoais** (LGPD – Lei nº 13.709/2018), o ATPBR recebe exclusivamente dados previamente **desidentificados** pelas instituições participantes. A plataforma não coleta informações de identificação direta nem determinadas variáveis que possam aumentar o risco de reidentificação dos participantes, e as chaves que associam os códigos às pessoas permanecem na instituição de origem. Como os registros de um mesmo participante continuam ligados entre si, esses dados são **pseudonimizados**, e não anônimos: por isso, recebem as proteções que a LGPD prevê para dados pessoais sensíveis. Os dados são armazenados e processados seguindo práticas de

segurança da informação e governança de dados adequadas à pesquisa científica. Esses princípios são retomados ao longo da documentação e detalhados em **Governança de dados** (cap. 4) e **Políticas de uso** (cap. 13).

Dessa forma, o ATPBR busca conciliar a proteção da privacidade dos participantes com a **promoção de uma pesquisa científica aberta**, **transparente** e **reprodutível**.

Saiba mais

- Site oficial: www.atpbr.com.br
- Contato: atpbr@inca.gov.br
- Artigo de referência: Boroni M, Costa Conteville L, Ribeiro Pinto LF, P B Viola J. *The Brazilian Population Tumor Atlas Project: addressing representation in cancer genomics in Brazil*. Lancet Reg Health Am. 2026;56:101397. doi:10.1016/j.lana.2026.101397

CAPÍTULO 02

Comece aqui

Bem-vindo ao ATPBR!

Esta seção foi criada para ajudar novos usuários a compreender rapidamente a estrutura da plataforma e encontrar os recursos necessários para suas atividades. Independentemente do seu perfil, recomendamos a leitura deste capítulo antes dos demais.

Como a documentação está organizada

Os capítulos seguem o **ciclo de vida dos dados** no ATPBR, da submissão à exploração dos resultados. Cada capítulo parte do anterior e indica, ao final, o próximo passo:

1. **Sobre o ATPBR**: objetivos, missão e princípios do projeto.
2. **Comece aqui**: este capítulo.
3. **Arquitetura da plataforma**: visão geral do caminho que os dados percorrem.
4. **Governança de dados**: quem é responsável por cada etapa, como os participantes são protegidos e quais autorizações cada uso exige.
5. **Submissão de dados**: o passo a passo para cadastrar um projeto e enviar seus dados.
6. **Acesso aos dados**: o que é aberto, o que é controlado e como solicitar acesso.
7. **Modelo de dados**: como as informações cadastradas se organizam e se relacionam.
8. **Dicionário de dados**: os campos de cada módulo e as regras de preenchimento.
9. **Sistema de identificação e barcode**: como a plataforma identifica pacientes, amostras e ensaios.
10. **Tipos de dados**: os formatos aceitos na submissão e os arquivos gerados pela plataforma.
11. **Plataforma analítica**: como os dados enviados são processados.
12. **Navegador de resultados**: como explorar os resultados no cBioPortal do ATPBR.
13. **Políticas de uso**: responsabilidades de quem utiliza os dados.
14. **Perguntas frequentes**: respostas rápidas para as dúvidas mais comuns.

Escolha o seu perfil

Se você não pretende ler a documentação na ordem, comece pelo caminho mais próximo da sua necessidade:

Quero submeter dados

Leia **Governança de dados**, que explica as responsabilidades de quem envia dados e a documentação ética exigida. Depois, siga de **Submissão de dados** até **Tipos de dados**. Esses capítulos cobrem tudo o que é preciso para o cadastro.

Quero explorar resultados

Comece pelo **Navegador de resultados**. Para entender como esses resultados foram gerados, consulte a **Plataforma analítica**.

Quero acessar dados controlados

Leia **Governança de dados**, **Acesso aos dados** e **Políticas de uso**.

Precisa de ajuda?

Consulte a seção de **Perguntas frequentes**.
Se precisar de suporte adicional, entre em contato com a equipe do ATPBR pelo e-mail atpbr@inca.gov.br.

CAPÍTULO 03

Arquitetura da plataforma

A arquitetura do ATPBR foi projetada para suportar todo o ciclo de vida dos dados oncológicos, desde sua submissão até sua análise e disponibilização de resultados para a comunidade científica. A plataforma é composta por módulos interoperáveis que atuam de forma integrada para garantir rastreabilidade, padronização, escalabilidade e reprodutibilidade.

Este capítulo apresenta a **visão de conjunto**. Cada etapa descrita aqui é detalhada em um capítulo próprio, indicado ao longo do texto.

Visão Geral da Arquitetura

De forma simplificada, o fluxo de dados no ATPBR segue as seguintes etapas:



Em todas as etapas, os princípios de proteção dos participantes apresentados em **Sobre o ATPBR** (cap. 1) são aplicados por meio de controle de acesso por perfis e permissões, registro e auditoria de atividades e gestão das solicitações de acesso a dados controlados. As regras e os responsáveis por cada etapa estão em **Governança de dados** (cap. 4).

Ingestão de Dados

A ingestão é o recebimento dos metadados e dos arquivos de sequenciamento enviados pelos usuários. O processo, passo a passo, está em **Submissão de dados** (cap. 5). Os dados são organizados segundo o **Modelo de dados** (cap. 7) e recebem identificadores padronizados, descritos em **Sistema de identificação e barcode** (cap. 9).

Validação

Após a ingestão, os dados passam por verificações automáticas e manuais de integridade, formato, campos obrigatórios e consistência entre metadados e arquivos, além da conferência da documentação ética e das permissões de uso. Ser aceito pelo sistema não equivale a ser admitido: enquanto houver pendências, o conjunto fica bloqueado para divulgação e para novos usos. As verificações e a correção de erros estão descritas em **Submissão de dados** (cap. 5), e as regras de admissão, em **Governança de dados** (cap. 4).

Armazenamento de Dados

A arquitetura separa o armazenamento em **duas camadas complementares**, ligadas por um catálogo de metadados:

- **Camada de dados brutos:** guarda os arquivos de grande volume (os FASTQ submetidos e os CRAM gerados no processamento), preservando os originais para rastreabilidade e futuras reanálises.
- **Camada de dados estruturados:** guarda os dados processados e organizados para consulta, como dados clínicos harmonizados, alterações genômicas e matrizes de expressão gênica.
- **Metadados e catalogação:** o catálogo registra projetos, investigações, estudos, pacientes, amostras, ensaios, experimentos, arquivos e processamentos executados. É ele que permite localizar um conjunto de dados, compreender sua origem e reconstruir o histórico completo de geração das informações.

Processamento de Dados

Os FASTQ validados são processados por pipelines padronizados, que identificam variantes somáticas e quantificam a expressão gênica. Os pipelines, as versões e os arquivos gerados estão em **Plataforma analítica** (cap. 11).

Harmonização e Integração

Os resultados do processamento são padronizados, anotados e integrados aos dados clínicos, o que permite comparar estudos diferentes. Os dados harmonizados não sensíveis são carregados no cBioPortal do ATPBR.

Exploração e Compartilhamento

Os resultados podem ser explorados no **Navegador de resultados** (cap. 12) (cBioPortal do ATPBR). Os estudos e conjuntos de dados dos projetos podem ser consultados no Catálogo de Dados da plataforma, desde que sua disponibilização tenha sido autorizada pelos respectivos responsáveis. Para saber mais sobre os critérios de disponibilização, os níveis de acesso (aberto ou controlado) e os procedimentos para solicitar acesso aos dados, consulte a seção **Acesso aos dados** (cap. 6).

CAPÍTULO 04

Governança de dados

Em Arquitetura da plataforma (cap. 3), vimos o caminho técnico que os dados percorrem no ATPBR. Este capítulo trata das **regras** que valem ao longo desse caminho: quem é responsável por cada etapa, que condições um conjunto de dados precisa cumprir para entrar na plataforma e ser compartilhado, e como os participantes das pesquisas são protegidos.

Leia este capítulo **antes de submeter dados ou de solicitar acesso a dados controlados**. Os capítulos seguintes aplicam estas regras a cada situação: Submissão de dados (cap. 5) a quem envia dados, Acesso aos dados (cap. 6) a quem solicita acesso e Políticas de uso (cap. 13) a quem recebe os dados.

Documentos e instrumentos de governança do ATPBR

A governança do Atlas Tumoral da População Brasileira (ATPBR) se apoia em políticas institucionais, instrumentos de proteção dos participantes e termos que formalizam responsabilidades. Juntos, cobrem todo o ciclo de vida dos dados, da submissão à reutilização, em conformidade com a legislação brasileira e as boas práticas internacionais. A coluna Status indica se cada documento está aprovado, em revisão ou em elaboração; documentos que ainda não foram aprovados **não devem ser lidos como normas vigentes**.

Políticas institucionais

Documento	O que estabelece	A quem se destina	Status
Política de Governança de Dados	Princípios gerais de governança; papel do INCA como instituição coordenadora e mantenedora; instâncias de governança (CGN, CCP, CAD, CESP e SAB) e responsabilidades das instituições participantes; constituição e gestão do banco de dados; proteção dos participantes; segurança da informação e mecanismos de supervisão.	Todas as instituições, pesquisadores e profissionais envolvidos no ATPBR.	Em revisão. Será publicada na íntegra após aprovação.
Política de Fluxo e Acesso a Dados	Procedimentos de submissão, verificação, classificação e disponibilização dos dados; modalidades de acesso (público, registrado, controlado por visita aos dados e interno); critérios para solicitação e autorização de acesso, incluindo a pré-autorização por códigos DUO; atribuições do Comitê de Acesso a Dados; monitoramento e encerramento dos acessos concedidos.	Pesquisadores remetentes, solicitantes de acesso, Comitê de Acesso a Dados e operadores da plataforma.	Em revisão. Será publicada na íntegra após aprovação.

Proteção de dados e direitos dos participantes

Documento	O que estabelece	A quem se destina	Status
Relatório de Impacto à Proteção de Dados Pessoais (RIPD)	Identifica e avalia os riscos do tratamento de dados pessoais e sensíveis na plataforma, individuais e coletivos, e descreve as medidas de segurança, salvaguardas e mecanismos adotados para preveni-los ou mitigá-los, em conformidade com a LGPD.	Instituições responsáveis pelo tratamento dos dados, encarregados de proteção de dados, equipes de governança e de segurança da informação e órgãos de fiscalização competentes.	Em revisão. Documento interno, por conter informações sobre riscos e controles de segurança; será publicado um resumo institucional.
Modelo de Termo de Consentimento Livre e Esclarecido (TCLE) em camadas	Modelo de referência que apresenta, de forma clara e estruturada, as finalidades de coleta, armazenamento, compartilhamento e reutilização dos dados em pesquisas futuras, com escolhas por camada (uso futuro, compartilhamento internacional, uso comercial, recontato e retorno de achados) registradas como códigos DUO. **Não é uma autorização geral:** cada estudo deve avaliar a compatibilidade entre seu consentimento, sua aprovação ética e os usos pretendidos.	Pesquisadores responsáveis pela coleta de dados, instituições participantes, Comitês de Ética em Pesquisa e participantes das pesquisas.	Em elaboração. O modelo será publicado; cada estudo adota a versão aprovada pelo seu CEP.

Termos de responsabilidade e compartilhamento

Documento	O que estabelece	A quem se destina	Status
Termo de Acordo Institucional	Formaliza a adesão de cada instituição ao consórcio ATPBR e os acordos prévios exigidos pela Resolução CNS nº 738/2024 para bancos de dados constituídos com a participação de mais de uma instituição: contribuição, acesso e compartilhamento dos dados, responsabilidades, participação nos colegiados e crédito aos centros.	Instituições participantes e INCA.	Em elaboração. O modelo será publicado.
Termo de Responsabilidade do Pesquisador Remetente	Declarações e responsabilidades de quem submete dados ao ATPBR, aceitas na plataforma **antes da primeira submissão**: legitimidade da submissão, observância das aprovações éticas, compatibilidade com os consentimentos e autorizações aplicáveis, qualidade das informações e cumprimento das políticas da plataforma.	Pesquisadores responsáveis pela submissão de dados.	Em revisão. Será publicado na íntegra após aprovação.
Termo de Uso de Dados de Acesso Controlado (DAA)	Condições para o uso dos dados autorizados na instância segura do ATPBR, por visita aos dados e sem download: finalidade aprovada, confidencialidade, proibição de tentativas de reidentificação e de extração de dados individuais, exportação apenas de resultados verificados, medidas de segurança, comunicação de incidentes, custos de processamento quando aplicáveis e condições de encerramento do acesso.	Pesquisadores e instituições autorizados a acessar dados controlados.	Em elaboração. O modelo será publicado.
Acordo de Transferência de Dados (DTA)	Condições jurídicas, institucionais e técnicas para o envio de dados das instituições de origem ao ATPBR: finalidades autorizadas, responsabilidades, medidas de proteção, restrições de uso e condições de armazenamento e retenção.	Instituições de origem que contribuem com dados e INCA.	Em elaboração. O modelo será publicado.

As regras de funcionamento dos colegiados ficam em regimentos próprios, em elaboração: o **Regimento do CAD** (composição, prazos e recursos) e o **Regimento do CCP** (período de embargo, autoria e crédito aos centros).

Compromisso com o uso responsável dos dados

O ATPBR adota uma abordagem de acesso responsável, considerando a natureza dos dados, suas condições de uso e os direitos dos participantes das pesquisas.

As permissões de uso de cada conjunto de dados são classificadas com a **Data Use Ontology (DUO)**, da Global Alliance for Genomics and Health (GA4GH), o que padroniza as restrições de uso e sua interpretação nos pedidos de acesso. A classificação DUO auxilia a leitura das permissões, mas **não substitui a verificação do consentimento e da aprovação ética** de cada estudo.

O compartilhamento de dados observa as aprovações éticas, as bases legais aplicáveis e os instrumentos institucionais pertinentes, garantindo que cada conjunto seja utilizado exclusivamente dentro das condições autorizadas. Os dados de acesso controlado são analisados **por visita aos dados**, na instância segura do ATPBR, **sem download**.

Esses documentos seguem a legislação e as normas éticas aplicáveis à pesquisa com seres humanos, entre elas:

- Lei Geral de Proteção de Dados Pessoais (LGPD, Lei nº 13.709/2018).
- Lei nº 14.874/2024, que dispõe sobre a pesquisa com seres humanos, e sua regulamentação.
- Resoluções CNS nº 466/2012 (pesquisa com seres humanos), nº 340/2004 (genética humana) e nº 738/2024 (bancos de dados para pesquisa).
- Resolução CD/ANPD nº 19/2024 (transferência internacional de dados pessoais).
- Políticas de Governança, de Segurança Cibernética e de Gestão de Crises do INCA.

As políticas do ATPBR **complementam, mas não substituem**, a aprovação ética de cada pesquisa, os termos de consentimento e os acordos firmados entre as instituições. Os documentos aprovados são publicados na íntegra nesta documentação, para consulta antes do aceite do Termo de Responsabilidade ou de um pedido de acesso; os documentos internos, como o RIPD, têm resumo público. Dúvidas podem ser enviadas para atpbr@inca.gov.br.

Princípios

A gestão dos dados no ATPBR segue, entre outros, estes princípios:

- **Finalidade científica:** os dados são usados para objetivos de pesquisa legítimos, definidos e compatíveis com as condições em que foram obtidos.
- **Proteção dos participantes:** respeito à dignidade, à autonomia, à confidencialidade e às escolhas manifestadas por cada participante, e prevenção de riscos coletivos, como a estigmatização de grupos populacionais, étnicos, comunitários ou territoriais.
- **Necessidade e proporcionalidade:** cada uso recebe apenas os dados adequados à finalidade autorizada.
- **Regularidade ética:** as aprovações éticas, os termos de consentimento e as decisões éticas pertinentes são sempre observados.
- **Qualidade e integridade científica:** os dados são mantidos íntegros, contextualizados e rastreáveis até a sua origem.
- **Segurança e prevenção:** as salvaguardas são proporcionais aos riscos de acesso indevido, reidentificação e uso incompatível.
- **Segregação de funções:** quem prepara, quem valida, quem decide e quem executa são pessoas ou instâncias distintas.
- **Responsabilização:** as decisões ficam registradas, de modo que se possa demonstrar a regularidade de cada operação.
- **Transparência e interoperabilidade:** as regras da plataforma são públicas, e os dados seguem metadados, vocabulários controlados e identificadores padronizados.

Autorizações diferentes, que não se substituem

Um mesmo conjunto de dados pode depender de várias autorizações. Cada uma tem uma finalidade própria, e nenhuma dispensa as outras:

Autorização	Quem concede	O que garante
Aprovação ética	Comitê de Ética em Pesquisa (CEP) ou Comissão Nacional de Ética em Pesquisa (Conep)	Que a pesquisa, de origem ou nova, é eticamente regular
Consentimento do participante (ou sua dispensa)	O participante; a dispensa só pode ser concedida pela instância ética competente	Que o participante autorizou o uso dos seus dados, nas condições do termo assinado
Anuência da instituição provedora	Representante habilitado da instituição que forneceu os dados	Que a instituição de origem concorda com aquele compartilhamento
Autorização de acesso	Comitê de Acesso a Dados (CAD) do ATPBR	Que um pedido específico de acesso controlado é compatível com as condições do conjunto

Duas consequências práticas:

- A aprovação ética do estudo de origem **não autoriza**, por si só, qualquer novo uso ou compartilhamento dos dados.
- Ser aceito pelo sistema **não equivale** a ser admitido. O preenchimento dos campos e o envio dos arquivos não constituem, sozinhos, autorização de incorporação, publicação ou uso científico.

Quem é quem

A governança do consórcio se apoia em seis instâncias com papéis complementares: três decidem, duas aconselham e uma executa.



Participante	Principais responsabilidades
INCA – Instituição coordenadora e mantenedora	Desenvolveu o ATPBR e mantém sua infraestrutura, operação e equipe executiva (Núcleo Executivo/LBBC). As decisões sobre ciência, acesso e ética são tomadas pelos colegiados do consórcio (CGN, CCP, CAD, CESP e SAB), nos quais os centros participantes têm assento e voto. Para fins da LGPD, o INCA atua como **controlador** dos dados tratados na plataforma, em conjunto com os centros participantes, que permanecem controladores dos dados que originam.
Instituições de origem (provedoras)	Asseguram a regularidade da coleta e do consentimento; removem os identificadores diretos e guardam, separadamente, as **chaves de correspondência**; informam permissões e restrições; comunicam alterações e retiradas de consentimento; cooperam no atendimento aos participantes e na resposta a incidentes. Para fins da LGPD, permanecem controladoras dos dados que originam.
Pesquisadores remetentes	Preparam e enviam os dados conforme os padrões da plataforma, apresentam a documentação ética e as permissões de cada conjunto e aceitam o Termo de Responsabilidade do Pesquisador Remetente.
Núcleo Executivo e Curadoria Operacional (LBBC/INCA)	Equipe interna que executa a operação: verifica integridade, completude, qualidade, proveniência e documentação ética de cada conjunto, registra pendências, roda os pipelines, cura dados e metadados (incluindo os códigos DUO), apoia os centros na submissão, monitora e audita os acessos concedidos e secretaria os comitês, **sem voto**. Reporta ao CGN.
Equipes técnicas e de segurança	Implementam os controles da plataforma: perfis de acesso, autenticação, criptografia, registros de auditoria, cópias de segurança e tratamento de vulnerabilidades.
Conselho de Governança Nacional (CGN)	**Instância máxima e deliberativa**, com 9 membros com voto e 2 observadores (Decit/MS e CNPq). Define a estratégia, aprova a entrada de centros e as políticas gerais, resolve conflitos entre os comitês e decide, junto com o CAD, os casos excepcionais, como uso comercial e transferência internacional ampla.
Comitê Científico e de Publicações (CCP)	**Deliberativo**, com 9 membros. Avalia propostas de análise, define projetos prioritários, evita sobreposição entre grupos, aplica as regras de embargo e autoria e registra manuscritos e apresentações.
Comitê de Acesso a Dados (CAD)	**Deliberativo**, com 7 membros, entre eles um membro leigo indicado pelo CESP e um representante do Banco Nacional de Tumores (BNT), que articula o comitê com o BNT e com os dados clínico-patológicos. Define as regras de compatibilidade com os códigos DUO, decide **em até 30 dias** os pedidos de acesso controlado não pré-autorizados e pode **suspender ou revogar** acessos em caso de descumprimento.
Comitê de Ética, Sociedade e Participação (CESP)	**Consultivo**, com 5 a 7 membros, incluindo representantes de pacientes e de minorias étnico-raciais. Acompanha os riscos coletivos, os usos proibidos e a participação social, e indica o membro leigo do CAD.
Conselho Consultivo Científico Internacional (SAB)	**Consultivo**, com 5 membros internacionais. Orienta o alinhamento do ATPBR aos padrões e às boas práticas internacionais, como os da GA4GH.
Pesquisadores e instituições destinatários	Usam os dados apenas na finalidade autorizada, mantêm confidencialidade e segurança, **não tentam reidentificar participantes** e cumprem o instrumento de uso firmado; respeitam os períodos de embargo e citam o ATPBR e os centros de origem.

As funções são exercidas por pessoas formalmente designadas pelo INCA, com substitutos definidos, e organizadas em uma **matriz de responsabilidades** que separa quem fornece os dados, quem os verifica, quem decide sobre sua incorporação ou disponibilização e quem executa e supervisiona. Os membros dos

colegiados têm mandato de 2 anos, renovável uma vez, declaram **conflitos de interesse** todo ano e a cada pauta, e **quem submete um conjunto não vota sobre ele**. Nenhum membro ocupa cadeira no CCP e no CAD ao mesmo tempo, e não há cadeira para empresas.

O caminho dos dados e quem atua em cada etapa



As etiquetas abaixo de cada etapa indicam as instâncias que atuam nela.

Proteção dos participantes

Os dados recebidos pelo ATPBR **não incluem informações de identificação direta** dos participantes.

Cada participante e amostra recebe um código, permitindo integrar informações clínicas e genômicas sem identificar diretamente as pessoas. A correspondência entre os códigos e as identidades fica protegida, exclusivamente sob responsabilidade da **instituição de origem**.

Todos os dados são tratados como **informações pessoais sensíveis**, seguindo a Lei Geral de Proteção de Dados (LGPD), com controles de acesso e cuidados adicionais para os dados genéticos.

Regras de proteção dos dados

Antes do envio ao ATPBR, todas as instituições seguem regras padronizadas:

- Remoção dos **identificadores pessoais** (nome, CPF, Cartão Nacional de Saúde, RG, número de prontuário, matrícula, endereço, telefone e e-mail).
- Uso dos **códigos** de paciente e de amostra definidos pelo ATPBR (veja Sistema de identificação e barcode (cap. 9)).
- Substituição das datas de eventos clínicos por **número de dias** a partir do diagnóstico (veja Dicionário de dados (cap. 8)).
- Verificação de **campos livres, nomes de arquivos, cabeçalhos e metadados técnicos**, onde identificadores costumam aparecer sem que se perceba.

Ciclo de vida de um conjunto de dados

Etapa	O que acontece
1. Cadastro e envio	O pesquisador remetente aceita o Termo de Responsabilidade, cadastra Projeto, Investigação, Estudo, Pacientes, Amostras, Ensaio, Arquivos e Experimentos e informa a documentação ética e as permissões de uso (veja Submissão de dados (cap. 5)).
2. Verificação de admissibilidade	Três verificações distintas e complementares: (i) **automatizada**, de campos obrigatórios, estrutura e vocabulários; (ii) **técnica**, de integridade, coerência e correspondência entre pacientes, amostras, experimentos e arquivos; e (iii) **ético-documental**, de pareceres, consentimento e permissões de origem.
3. Resultado da validação	A equipe de validação registra o conjunto como **apto**, **pendente de complementação** ou **não admitido**, com a motivação e o responsável pela decisão.
4. Processamento e curadoria	Os dados aptos são processados pela Plataforma analítica (cap. 11). A proveniência é preservada, e as restrições de origem acompanham os dados derivados.
5. Classificação e divulgação	Cada material é classificado como de acesso público, controlado ou interno, e o que for público passa por uma avaliação de risco de reidentificação antes de ser divulgado (veja Acesso aos dados (cap. 6)). Antes da abertura, o conjunto pode cumprir um **período de embargo** (moratória de publicação), durante o qual a primeira publicação é reservada ao projeto de origem; as regras de embargo e autoria são aplicadas pelo Comitê Científico e de Publicações (CCP).
6. Acesso controlado	Pesquisadores solicitam acesso, e o CAD decide caso a caso (veja Acesso aos dados (cap. 6)).
7. Conservação e descarte	A necessidade de conservar dados e documentos é revista periodicamente, conforme as finalidades científicas, as obrigações legais e éticas e as condições de origem.

Conjuntos **pendentes** ou **não admitidos permanecem bloqueados**: não aparecem para terceiros, não são publicados e não podem ser usados em novas pesquisas até a regularização. Mudanças relevantes depois da validação, como um novo documento, uma nova restrição ou novos pacientes, levam o conjunto a uma **nova verificação**.

Documentação ética e permissões de cada conjunto

Cada conjunto de dados fica vinculado aos documentos que comprovam que ele pode entrar na plataforma e ser usado. Esse conjunto de registros forma o **dossiê** do conjunto.

Nível	O que é informado e conferido
Projeto	Título, responsável científico, instituições participantes e aprovação ética, com os documentos comprobatórios, sua versão e sua validade
Estudo	Parecer ético pertinente, características da coorte, condições de divulgação no catálogo e restrições do estudo
Pacientes e Amostras	Tipo de consentimento, permissões e restrições de uso e códigos de identificação
Ensaio, Experimentos e Arquivos	As condições de uso acompanham os dados produzidos

Os documentos esperados incluem o parecer do CEP ou da Conep, a versão do TCLE utilizada (ou a decisão de dispensa de consentimento), o instrumento de transferência de dados (DTA) e o registro das permissões de uso. **Um campo preenchido sem o documento correspondente não supre a exigência de comprovação.**

As permissões registradas em cada nível (Estudo, Paciente, Amostra, Ensaio e Arquivo) são consideradas **em conjunto**. **Prevalece sempre a restrição mais protetiva**, e uma divergência entre níveis, por exemplo um estudo liberado para uso geral com um paciente que recusou o uso futuro, **impede a liberação** até ser resolvida e documentada.

As permissões e restrições seguem a **Data Use Ontology (DUO)** da GA4GH, que padroniza condições como uso geral em pesquisa (GRU), pesquisa em saúde (HMB), doença específica (DS), uso não comercial (NCU), exigência de aprovação ética (IRB), colaboração com o pesquisador provedor (COL), restrição geográfica (GS) e moratória de publicação (MOR).

Uso futuro dos dados

Usar os dados em uma nova pesquisa exige respeitar as condições do estudo de origem e as escolhas de cada participante:

Situação do consentimento original	O que é necessário
O uso futuro foi autorizado	A nova pesquisa pode usar os dados nas condições autorizadas, com sua própria avaliação ética
O uso futuro não foi previsto	Novo consentimento ou dispensa concedida pela instância ética competente
O uso futuro foi expressamente recusado	Somente com novo consentimento do participante ou do seu responsável legal

O CAD não concede dispensa de consentimento: essa competência é exclusiva da instância ética.

Acesso aos dados e Comitê de Acesso a Dados

O que é público, o que é controlado e o que é de uso interno, assim como as etapas de um pedido de acesso controlado e a forma como o CAD decide, estão em Acesso aos dados (cap. 6). Dois pontos valem para todos os casos:

- Atuar no desenvolvimento do software ou na manutenção da infraestrutura **não dá acesso** aos dados de pesquisa. As permissões internas seguem uma matriz de perfis e o princípio do menor privilégio.
- O CAD, cuja composição está em Quem é quem, reúne-se mensalmente, e seus membros não participam das deliberações em que estejam impedidos. Quórum e demais regras estão no **Regimento do CAD**.

Amostras biológicas (fora do escopo do ATPBR)

O ATPBR **gerencia apenas dados**. A governança das amostras biológicas (coleta, guarda, acesso, uso e transferência) é de competência do Banco Nacional de Tumores (BNT) do INCA e das instituições que as guardam, e não é regulada pelas políticas do ATPBR; o mesmo vale para os instrumentos sobre amostras, como o Acordo de Transferência de Material Biológico (MTA). A autorização de acesso a dados **não inclui acesso às amostras** correspondentes.

As permissões e restrições que se aplicam a uma amostra **continuam valendo para os dados produzidos a partir dela**.

Direitos dos participantes e retirada de consentimento

Os participantes exercem seus direitos (acesso, correção, restrição e retirada dos dados) por meio da **instituição de origem**, que é quem guarda as chaves de correspondência. O atendimento é coordenado com o encarregado pelo tratamento de dados pessoais do INCA, e a plataforma localiza os registros **pelos códigos**, sem precisar receber nome ou documento do participante.

Quando há retirada de consentimento, nova restrição ou outra solicitação do participante:

- os conjuntos afetados são identificados pelos códigos de origem;
- novos usos incompatíveis são bloqueados;
- os pesquisadores que já receberam os dados são comunicados, quando necessário;
- as providências são documentadas, ressalvadas as hipóteses legais de retenção.

A instituição de origem e o pesquisador remetente devem **comunicar prontamente** ao ATPBR qualquer alteração que afete os dados enviados, como mudança de parecer ético, retirada de consentimento, nova restrição ou incidente de segurança.

Segurança, incidentes e rastreabilidade

A plataforma segue a Política de Segurança Cibernética do INCA, com controles adicionais proporcionais aos seus riscos: gestão de perfis, acesso limitado ao necessário para cada função, autenticação reforçada para acessos privilegiados, criptografia, cópias de segurança e continuidade das operações.

São mantidos registros das solicitações, decisões, anuências, habilitações, acessos, downloads, suspensões, revogações e encerramentos, **protegidos contra alteração indevida**. Esses registros permitem reconstruir quem acessou qual conjunto, quando e com base em qual autorização.

Incidentes de segurança são tratados pelos procedimentos de privacidade e segurança do INCA, com a participação da equipe do ATPBR e a cooperação das instituições de origem. Pesquisadores que suspeitem de incidente, reidentificação ou uso indevido devem **comunicar o INCA imediatamente**.

Os riscos das operações da plataforma são avaliados no **Relatório de Impacto à Proteção de Dados Pessoais (RIPD)** do ATPBR.

Monitoramento e revisão

A coordenação do ATPBR acompanha indicadores sobre a regularidade dos conjuntos, as decisões de disponibilização, as solicitações dos participantes, os incidentes e as medidas corretivas.

Mudanças relevantes na arquitetura da plataforma, nas categorias de dados, nas funcionalidades de acesso, nas instituições participantes ou nas condições de uso levam a uma avaliação do impacto sobre a governança, o protocolo ético, os instrumentos jurídicos e os controles de segurança.

A implementação dos controles é progressiva, mas **cada operação só começa depois que os controles necessários a ela estão em funcionamento**. Por isso, a divulgação pública e o acesso controlado só são ativados depois de concluídos os seus respectivos procedimentos.

PARTE 2

Enviar e compartilhar dados

Como cadastrar projetos e dados na plataforma e como eles podem ser compartilhados.

05 Submissão de dados

06 Acesso aos dados

CAPÍTULO 05

Submissão de dados

A submissão de dados é o processo pelo qual centros de pesquisa, hospitais, laboratórios e grupos de pesquisa incorporam projetos e seus respectivos dados ao ATPBR. É a etapa de **ingestão** apresentada em Arquitetura da plataforma (cap. 3): a partir dela, os dados seguem para validação, processamento e compartilhamento.

Este capítulo descreve **o que fazer e em que ordem**. Depois de Acesso aos dados (cap. 6), que explica como os dados submetidos são compartilhados, os capítulos seguintes detalham como as informações se organizam (Modelo de dados (cap. 7)), como cada campo deve ser preenchido (Dicionário de dados (cap. 8)), como os identificadores são gerados (Sistema de identificação e barcode (cap. 9)) e quais formatos são aceitos (Tipos de dados (cap. 10)).

Quem Pode Submeter Dados

A submissão de dados pode ser realizada por pesquisadores, instituições e centros de pesquisa que desejem incorporar estudos ao ATPBR, desde que atendam aos requisitos técnicos, éticos e de governança estabelecidos pela plataforma (veja os critérios mínimos em Sobre o ATPBR (cap. 1) e as regras completas em Governança de dados (cap. 4)).

Antes de enviar dados, o pesquisador remetente aceita o **Termo de Responsabilidade do Pesquisador Remetente**, apresentado na área de submissão. Ao aceitá-lo, ele declara, entre outros pontos, que:

- os dados são autênticos e as informações prestadas são verdadeiras, completas e atualizadas;
- os dados foram desidentificados conforme os padrões do ATPBR, e as chaves que ligam os códigos aos participantes permanecem separadas e protegidas na instituição de origem;
- vai informar as aprovações éticas, a versão do TCLE (ou a dispensa de consentimento) e as permissões e restrições de uso de cada conjunto, e apresentar os documentos comprobatórios;
- vai comunicar ao ATPBR mudanças nas aprovações éticas, retiradas de consentimento, novas restrições e incidentes de segurança de que tiver conhecimento.

Tenha também em mãos a documentação ética do projeto e dos estudos: o parecer do CEP ou da Conep, a versão do TCLE utilizada ou a decisão de dispensa de consentimento, o instrumento de transferência de dados (DTA), quando houver, e as permissões e restrições de uso dos dados. A equipe do ATPBR pode solicitar esses documentos durante a validação.

Passo a Passo

A submissão é feita pelo portal da plataforma. Cada etapa depende da anterior, e o painel do projeto indica essa ordem com números e setas.

Etapa	O que cadastrar	Como
1	**Projeto**, incluindo o acrônimo e as instituições associadas	Formulário
2	**Investigação** (linha de pesquisa)	Formulário
3	**Estudo** (um tipo de câncer)	Formulário
4	**Pacientes** do Estudo	Formulário ou planilha-modelo
5	**Amostras** de cada paciente	Formulário ou planilha-modelo
6	**Ensaio** (uma corrida de sequenciamento)	Formulário
7	**Arquivos de dados** (FASTQ) do Ensaio	Envio pelo portal **ou** caminho na Plataforma Multiusuário de Bioinformática do INCA
8	**Experimentos** (metadados técnicos de cada amostra sequenciada)	Formulário ou planilha-modelo

Alguns pontos de atenção em cada etapa:

- **Projeto:** as instituições associadas informadas aqui são as que aparecerão depois como opção de **centro coletor** no cadastro de pacientes.
- **Estudo:** dentro de um mesmo projeto, cada tipo de câncer só pode ter um Estudo.
- **Pacientes e Amostras:** uma amostra só pode ser cadastrada para um paciente já registrado. Os eventos clínicos são informados em **número de dias** a partir do diagnóstico, e não como datas (veja Dicionário de dados (cap. 8)).
- **Ensaio:** cada corrida de sequenciamento é um Ensaio separado. No cadastro, também são indicadas as **condições de compartilhamento** desejadas para o Ensaio. Essa indicação não publica os dados por si só: a divulgação depende da validação do conjunto e de aprovação própria (veja Acesso aos dados (cap. 6)).
- **Arquivos de dados:** os FASTQ precisam estar disponíveis **antes** do cadastro dos Experimentos. Há duas formas de fazer isso:
  - **Envio pelo portal:** na área *Arquivos de Dados*, selecione o Ensaio e envie os arquivos FASTQ. Cada arquivo fica vinculado ao Ensaio e herda suas permissões de compartilhamento.
  - **Arquivos já armazenados no INCA:** se os FASTQ já estiverem no repositório da **Plataforma Multiusuário de Bioinformática do INCA**, não é necessário enviá-los novamente. Basta informar o **caminho absoluto do diretório** onde estão os arquivos no campo *Repositórios Associados* do Ensaio.
- **Experimentos:** um experimento só pode ser cadastrado para uma amostra já registrada. Os campos `fastq_1` e `fastq_2` devem conter o **nome exato** dos arquivos FASTQ do Ensaio. A plataforma avisa quando um nome informado não corresponde a nenhum arquivo enviado.

Antes de Submeter

Recomenda-se verificar se:

- Os arquivos FASTQ estão completos e íntegros, e no formato aceito (veja **Tipos de dados** (cap. 10)).
- Os campos obrigatórios foram preenchidos e os vocabulários controlados foram utilizados (veja **Dicionário de dados** (cap. 8)).
- Os códigos de paciente e de amostra seguem as regras de desidentificação (veja **Sistema de identificação e barcode** (cap. 9)).
- Não há informações pessoais ou identificadores diretos de participantes (nome, CPF, Cartão Nacional de Saúde, número de prontuário, contatos) nos metadados, **nos campos de texto livre, como descrições e observações, nos nomes dos arquivos nem nos cabeçalhos dos arquivos**.
- As datas de eventos clínicos foram convertidas em número de dias a partir do diagnóstico.
- A documentação ética e as permissões de uso de cada conjunto estão disponíveis (veja **Governança de dados** (cap. 4)).
- As planilhas foram preenchidas a partir dos templates oficiais, sem alterar sua estrutura.

O cumprimento desses requisitos reduz a ocorrência de erros e acelera o processo de validação.

Validação dos Dados

Após o envio, cada conjunto passa por três verificações distintas e complementares, feitas por responsáveis diferentes:

- **Verificação automatizada:** campos obrigatórios, estrutura e uso dos vocabulários controlados.
- **Verificação técnica:** formato e integridade dos arquivos, consistência entre pacientes, amostras, experimentos e arquivos e compatibilidade com o modelo de dados do ATPBR.
- **Verificação ético-documental:** correspondência entre o estudo cadastrado, o protocolo e a aprovação ética, as condições de consentimento ou dispensa e as permissões e restrições de uso.

No cadastro por planilha, uma prévia mostra as linhas lidas e os erros encontrados antes de salvar, para que você possa conferir e confirmar.

Ao final, a equipe de validação registra o conjunto como **apto**, **pendente de complementação** ou **não admitido**, com a motivação. O fato de os dados terem sido aceitos pelo sistema **não significa** que foram admitidos: conjuntos pendentes ou não admitidos ficam bloqueados para divulgação e para novos usos até a regularização.

Correção de Erros

Caso sejam detectados problemas durante a validação, o sistema apresentará mensagens indicando os erros encontrados e as ações necessárias para correção.

Dependendo do tipo de inconsistência, poderá ser necessário:

- Atualizar metadados.
- Corrigir nomenclaturas.

- Reenviar arquivos.
- Completar informações obrigatórias.
- Ajustar relacionamentos entre amostras e experimentos.
- Apresentar documentos éticos ou esclarecer permissões de uso.

Após as correções, os dados poderão ser reenviados para uma nova validação. Submissões que não atendam aos requisitos desta documentação poderão ser consideradas não admitidas.

Se, depois do envio, houver alguma mudança que afete os dados (novo parecer ético, retirada de consentimento de um participante, nova restrição de uso ou incidente de segurança), comunique a equipe do ATPBR pelo e-mail atpbr@inca.gov.br. Mudanças relevantes levam o conjunto a uma nova verificação.

CAPÍTULO 06

Acesso aos dados

Ao submeter dados ao ATPBR (veja **Submissão de dados** (cap. 5)), você indica como gostaria que eles fossem compartilhados. Este capítulo explica as modalidades de acesso, que equilibram o compartilhamento científico, a privacidade dos participantes e os requisitos éticos e regulatórios, e descreve como solicitar acesso a dados controlados. Os princípios e responsabilidades por trás dessas regras estão em **Governança de dados** (cap. 4).

Modalidades de acesso

Modalidade	O que inclui	Quem acessa
Público	Metadados descritivos de estudos e ensaios, informações do catálogo e resultados aprovados para divulgação, incluindo mutações somáticas e um conjunto clínico mínimo com variáveis agrupadas (idade em faixas, sobrevida arredondada, raça/cor agrupada)	Qualquer pessoa
Registrado	Conjuntos de menor risco, classificados para essa modalidade segundo critérios aprovados pelo CAD	Pesquisadores com identidade e vínculo institucional verificados, após aceite do termo padrão de uso; cada acesso é registrado, sem deliberação individual do CAD
Controlado	Arquivos de sequenciamento, resultados intermediários, variantes germinativas, dados genômicos individualizados, matrizes ômicas completas e dados clínicos que exijam proteção adicional	Pesquisadores autorizados pelo Comitê de Acesso a Dados, nas condições aprovadas, ou por pré-autorização, quando a finalidade é integralmente compatível com os códigos DUO de todos os conjuntos (regras definidas pelo CAD).
Interno operacional	Conjuntos em validação ou em processamento e informações administrativas	Equipes autorizadas do ATPBR, conforme suas funções

Quem decide o que é compartilhado

No cadastro de cada **Ensaio**, o responsável pelo projeto indica se aquele ensaio pode ser disponibilizado no ATPBR (veja **Submissão de dados** (cap. 5)). Essa indicação é o ponto de partida, mas não basta: a divulgação só acontece depois que o conjunto é validado como **apto** e que a publicação é aprovada, com uma avaliação dos riscos de reidentificação. As permissões registradas no Estudo, nos Pacientes, nas Amostras, no Ensaio e nos Arquivos são consideradas em conjunto, e prevalece sempre a mais restritiva.

Quando o consentimento dos participantes, as permissões de origem e as políticas do ATPBR permitem:

- Os **metadados** do ensaio (título, descrição e tipo de ômica) são incluídos no **catálogo de dados** da plataforma.
- Os **resultados das análises** aprovados para divulgação podem ser publicados no cBioPortal do ATPBR.
- Os **arquivos FASTQ e CRAM** permanecem sob acesso controlado.

Dados de acesso público

Os resultados aprovados para divulgação pública no cBioPortal do ATPBR (veja **Navegador de resultados** (cap. 12)) podem ser consultados e baixados sem necessidade de autorização adicional.

O **catálogo de dados** da plataforma lista os estudos e conjuntos de dados de todos os projetos, com o tipo de câncer, o tipo de ômica e o número de pacientes, amostras e experimentos. Para cada conjunto, o catálogo indica se você já tem acesso (por fazer parte do projeto ou por concessão) ou se é necessária autorização.

Toda divulgação pública segue estas regras:

- Antes da publicação, avaliam-se coortes pequenas, atributos raros, combinações de filtros clínicos, demográficos e genômicos e consultas sucessivas que permitam reconstruir informações individuais.
- Quando o risco não é aceitável, as informações são agregadas, apresentadas em faixas ou omitidas, ou o conjunto passa ao acesso controlado. Por padrão, no portal público a idade ao diagnóstico aparece em faixas de 5 anos, a sobrevida em meses arredondados, a raça/cor em categorias agrupadas, e contagens abaixo do limiar mínimo são suprimidas. Variantes germinativas, leituras brutas e datas nunca são públicas. Os limiares e técnicas são definidos pela coordenação do ATPBR, com validação técnica.
- Cada aprovação registra o responsável, a versão do conjunto, os parâmetros de publicação, os testes realizados e a data. Mudanças relevantes nos dados, nos filtros ou nos indicadores exigem nova avaliação.
- Aprovar a divulgação do registro descritivo **não** libera os arquivos de sequenciamento, os registros clínicos individuais nem as matrizes ômicas associadas.

Dados de acesso controlado

Alguns conjuntos de dados possuem restrições de acesso devido ao seu potencial de identificação, à sua sensibilidade ou aos termos de consentimento. Nessa categoria estão incluídos, entre outros:

- Arquivos de sequenciamento bruto ou alinhado (FASTQ e CRAM).

- Resultados intermediários do processamento.
- Variantes germinativas e demais dados genômicos individualizados.
- Matrizes ômicas completas ou não aprovadas para divulgação (as matrizes processadas e aprovadas aparecem no Navegador de resultados).
- Dados clínicos individuais além das variáveis publicadas no Navegador de resultados.
- Informações clínicas cuja granularidade ou combinação de atributos exija proteção adicional.
- Outros dados definidos pelas políticas de governança do ATPBR.

O acesso a esses recursos exige uma solicitação formal, analisada pelo **Comitê de Acesso a Dados (CAD)** do ATPBR. A autorização do CAD não substitui a aprovação ética da nova pesquisa nem a anuência da instituição que forneceu os dados, quando exigida.

Acesso interno operacional

Conjuntos em validação ou em processamento e informações administrativas são acessados apenas pelas equipes autorizadas do ATPBR, conforme suas funções e o princípio do menor privilégio. Atuar no desenvolvimento do software ou na manutenção da infraestrutura não dá acesso aos dados de pesquisa. Acessos privilegiados ou excepcionais dependem de justificativa, autorização, prazo limitado e registro, e são revistos depois.

Como solicitar acesso a dados controlados

Pesquisadores interessados em acessar dados de acesso controlado devem registrar a solicitação no módulo de pedidos de acesso da plataforma ATPBR, após login com conta institucional. Dúvidas sobre o processo podem ser enviadas para atpbr@inca.gov.br. A equipe fornecerá as orientações sobre o processo de solicitação e a documentação necessária, de acordo com a natureza dos dados e as políticas de acesso vigentes.

O que o pedido deve conter

- Identificação e contato institucional do pesquisador responsável e da sua instituição.
- Título, objetivos, justificativa científica, metodologia e duração prevista da pesquisa.
- Os conjuntos, variáveis, arquivos e o nível de detalhe solicitados, com a justificativa da necessidade de cada categoria de dado.
- O protocolo e a aprovação ética da nova pesquisa, quando exigível, e a decisão de dispensa de consentimento, se for o caso.
- A equipe que terá acesso, o ambiente de tratamento e armazenamento, as medidas de segurança, o país de acesso remoto, quando houver, e a possibilidade de cópia ou extração.
- O compromisso de confidencialidade, de não reidentificação, de não retransmissão e de respeito às condições de origem.

Cada colaborador precisa ser identificado individualmente, estar vinculado ao projeto e aceitar as mesmas obrigações. O simples cadastro de usuário na plataforma não dá acesso a materiais restritos.

Etapas do pedido

Etapa	Responsável	Resultado
1. Solicitação	Pesquisador solicitante	Pedido registrado, com número de identificação, ou devolvido para completar
2. Triagem	Equipe designada pelo ATPBR	Pedido instruído, diligência (pedido de complementação) ou bloqueio
3. Compatibilidade	Equipe de validação e coordenação	Parecer sobre a compatibilidade do pedido com o TCLE, o protocolo e as permissões de cada conjunto
4. Anuência da origem	Representante autorizado da instituição provedora	Manifestação registrada ou comprovação de anuência prévia
5. Deliberação	CAD	Decisão fundamentada, com condições e prazo
6. Formalização	Pesquisador, sua instituição e INCA	Instrumento de uso de dados firmado
7. Habilitação	Equipe técnica autorizada	Acesso ativado e registrado, por usuário
8. Monitoramento	Coordenação e equipe técnica	Revisão, manutenção ou suspensão do acesso
9. Encerramento	Equipe técnica e coordenação	Acesso desativado e destinação dos dados registrada

Como o CAD decide

O CAD confronta, **para cada conjunto**, a finalidade declarada com as permissões e restrições registradas na origem (veja **Governança de dados** (cap. 4)), sem estender a um conjunto as permissões de outro. A comparação automatizada de códigos DUO serve de apoio, mas não decide: casos inconclusivos, conflitos e pedidos excepcionais são sempre avaliados por pessoas.

O CAD pode deferir integralmente o pedido; deferir parcialmente, limitando conjuntos, variáveis, pessoas, finalidade ou prazo; deferir sob condições; determinar diligência para complementação; ou indeferir, com fundamentação. A decisão é comunicada ao solicitante por canal rastreável.

O instrumento de origem define se a instituição provedora se manifesta **caso a caso** ou se deu **anuência prévia** para usos compatíveis. No primeiro caso, a manifestação precisa ser expressa: o **silêncio não vale como anuência**. No segundo, verifica-se se o pedido está inteiramente coberto pelas permissões já formalizadas.

Depois da decisão

- Antes da liberação, o pesquisador e, quando aplicável, sua instituição firma o **instrumento de uso de dados** aprovado pelo INCA, que não pode ampliar as autorizações de origem.
- O acesso é **individual**, vinculado ao projeto e aos conjuntos autorizados, com prazo de expiração (em regra, 12 meses, com aviso antes do vencimento) e sem contas compartilhadas.
- A **renovação não é automática**: depende de nova comprovação das condições de elegibilidade.
- O acesso pode ser **suspenso preventivamente** diante de indícios de uso indevido, risco de reidentificação ou mudança nas permissões de origem, e **revogado** pelo CAD em caso de descumprimento.

- No **encerramento**, as credenciais e a instância de processamento são desativadas. Os dados permanecem no ATPBR; ficam com o pesquisador apenas os resultados cuja exportação foi verificada e autorizada.
- Pedidos de reconsideração podem ser apresentados diante de correção documental, fato novo ou erro material, mas não liberam acesso provisório.

O acesso a dados por pesquisador ou instituição no exterior depende, além das demais condições, das permissões de origem para compartilhamento internacional e de um mecanismo de transferência admitido pela LGPD.

As obrigações de quem recebe os dados estão em **Políticas de uso** (cap. 13).

PARTE 3

Organização dos dados

O modelo de dados, o dicionário, o sistema de identificação e os formatos aceitos.

07 Modelo de dados

08 Dicionário de dados

09 Sistema de identificação e barcode

10 Tipos de dados

CAPÍTULO 07

Modelo de dados

Em **Submissão de dados** (cap. 5), vimos a ordem em que as informações são cadastradas. O modelo de dados explica **por que** essa ordem existe: ele define como as informações são organizadas e relacionadas na plataforma, garantindo rastreabilidade, interoperabilidade e consistência ao longo de todo o ciclo de vida dos dados.

Visão Geral

O ATPBR utiliza uma estrutura hierárquica composta por quatro módulos principais:



Essa organização conecta o contexto científico da pesquisa aos pacientes, às amostras biológicas, aos procedimentos experimentais e aos arquivos de dados gerados.

Os módulos **Estudo** e **Ensaio** também contêm elementos específicos:

- **Estudo:** Pacientes e Amostras.
- **Ensaio:** Experimentos e Arquivos de dados.

Principais Módulos

Projeto

Representa o contexto científico geral da pesquisa e constitui o nível mais alto da hierarquia. Reúne uma ou mais linhas de pesquisa relacionadas e define os objetivos científicos, o escopo e as instituições participantes.

Investigação

Representa uma linha de pesquisa dentro de um Projeto. Agrupa estudos relacionados por um tema, objetivo científico ou hipótese, podendo abranger diferentes tipos de câncer.

Estudo

Representa uma unidade científica dedicada à caracterização de um único tipo de câncer dentro de uma Investigação. Define o desenho do estudo, as coortes, os critérios de seleção e as características clínicas dos participantes.

Contém dois elementos:

- **Pacientes:** participantes do estudo, cada um com um único registro na plataforma, independentemente do número de amostras associadas.
- **Amostras:** materiais biológicos coletados dos pacientes. Um paciente pode estar associado a várias amostras.

Ensaio

Representa uma corrida experimental utilizada para gerar dados ômicos a partir das amostras de um Estudo. Corridas distintas, mesmo quando utilizam a mesma abordagem ômica, devem ser cadastradas como Ensaios independentes. Um Estudo pode conter vários Ensaios.

Contém dois elementos:

- **Arquivos de dados:** arquivos brutos produzidos na corrida experimental, como FASTQ (veja **Tipos de dados**), utilizados como entrada para o processamento na plataforma.
- **Experimentos:** registros dos metadados técnicos associados a cada amostra processada na corrida, incluindo plataforma de sequenciamento, preparação da biblioteca e layout de leitura. Cada Experimento relaciona uma Amostra aos respectivos Arquivos de dados.

Como os Elementos se Relacionam



A figura apresenta a hierarquia dos módulos e as relações entre seus elementos.

Além de pertencer a um módulo, determinados elementos dependem de registros previamente cadastrados:

- Uma **Amostra** deve estar associada a um **Paciente**.
- Um **Experimento** deve estar associado a uma **Amostra** e aos respectivos **Arquivos de dados** disponíveis no Ensaio.

Essas dependências determinam a ordem de cadastro na plataforma e permitem rastrear os dados experimentais até a amostra e o paciente de origem.

Os campos de cada módulo e elemento estão descritos no capítulo seguinte. Alguns deles também compõem o identificador único de pacientes, amostras e ensaios, apresentado em **Sistema de identificação e barcode** (cap. 9).

CAPÍTULO 08

Dicionário de dados

Depois de conhecer os módulos e elementos do **Modelo de dados** (cap. 7), este capítulo descreve **o que preencher em cada um**. O dicionário de dados define a estrutura, o significado e as regras de preenchimento dos dados armazenados na plataforma, garantindo a consistência e a comparabilidade entre os dados submetidos pelos diferentes centros participantes.

Cada campo possui uma definição formal, tipo e formato esperados, obrigatoriedade de preenchimento e, quando aplicável, vocabulários controlados ou ontologias de referência. Alguns desses campos também compõem o **Sistema de identificação e barcode** (cap. 9) do ATPBR.

Organização do Dicionário de Dados

O dicionário de dados está organizado de acordo com os quatro módulos principais da plataforma:

- Projeto:
  - Campos obrigatórios:
    - Título do Projeto, Acrônimo, Descrição, Pesquisador(a) Principal (PI), Instituições Associadas, Aprovação Ética, Financiamento
  - Campos opcionais:
    - Co-PIs, Comitê Científico, Data Início, Data Fim
  - Campos associados a ontologias:
    - Instituições Associadas (associadas ao ROR ID)
  - Campos utilizados na composição do barcode:
    - Acrônimo (5 a 7 letras maiúsculas), Instituições Associadas (referência ao centro que coletou as amostras)
- Investigação:
  - Campos obrigatórios:
    - Título da Investigação
  - Campos opcionais:
    - Descrição
- Estudo:

  - Campos obrigatórios:
    - Título do Estudo, Investigação Associada, Tipo de Câncer, Descrição, Aprovação Ética Específica
  - Campos opcionais:
    - Publicações
  - Campos com ontologias de Referência:
    - Tipo de Câncer (associado ao OncoTree)
  - Dentro de um mesmo projeto, cada tipo de câncer só pode ter um Estudo.
- Ensaios:
  - Campos obrigatórios:
    - Título do Ensaio, Tipo de Ômica
  - Campos opcionais:
    - Descrição, Repositórios Associados
  - Campos utilizados na composição do barcode:
    - Tipo de Ômica
  - Campo para arquivos já armazenados no INCA:
    - Repositórios Associados: se os FASTQ já estiverem no repositório da Plataforma Multiusuário de Bioinformática do INCA, informe aqui o caminho absoluto do diretório (veja **Submissão de dados** (cap. 5)).
  - Ensaio também indica as condições de compartilhamento desejadas para os seus dados, que dependem de validação e aprovação (veja **Acesso aos dados** (cap. 6)).

De acordo com os três principais elementos do modelo de dados da plataforma de ingestão de dados, cada elemento possui um conjunto específico de atributos clínicos, biológicos e experimentais. A descrição completa de todos os atributos está disponível na planilha ATPBR_planilha_metadados.xlsx

Cada atributo descrito nesta planilha apresenta, quando aplicável:

- Nome do atributo
- Descrição
- Tipo de dado
- Obrigatoriedade
- Valores permitidos
- Vocabulário controlado ou ontologia

A seguir são apresentados apenas os principais campos obrigatórios de cada elemento. A descrição completa encontra-se na planilha de metadados.

- **Pacientes**
  - Principais campos obrigatórios:
    - `patient_id`: Código do participante, com 4 dígitos, definido pelo centro coletor e sem relação com dados pessoais
    - `collector_center`: Centro coletor (instituição), via vocabulário "Collector Center"

    - `disease_type`: Tipo de câncer do paciente, escolhido na classificação OncoTree entre os tipos do tecido definido no Estudo
    - `primary_site`: Órgão de origem do tumor
    - `consent_type`: Tipo de consentimento obtido para participação no estudo
    - `data_use_permission`: Permissão primária de uso dos dados/amostras desse paciente
    - `data_use_restriction`: Restrição adicional sobre o uso dos dados conforme a GA4GH Data Use Ontology (DUO). Um único valor; use "NRES" quando não houver restrição adicional
  - Campo opcional relacionado ao diagnóstico: `disease_subtype`: Subtipo OncoTree, abaixo do tipo de câncer, quando conhecido
  - Campo automático: `cid10_category`: Categoria CID-10 calculada pela plataforma a partir de `disease_type`, `disease_subtype` e `primary_site`. O usuário não preenche este campo.
  - Campos utilizados na composição do barcode: `patient_id`, `collector_center`, `cid10_category`
- **Amostras**
  - Principais campos obrigatórios:
    - `sample_id`: Código da amostra, com 4 dígitos, definido pelo centro coletor e sem relação com dados pessoais
    - `patient_id`: Identificador do Paciente já registrado ao qual esta amostra pertence
    - `sample_type`: Material biológico ou derivado biológico correspondente à amostra coletada.
    - `sample_status`: Classificação patológica da amostra no momento da coleta.
    - `sample_preservation`: Método utilizado para preservação da amostra após a coleta.
  - Campos utilizados na composição do barcode: `sample_id`, `sample_type`, `sample_status`, `sample_preservation`
  - Amostra exige o preenchimento de ao menos um sistema de estadiamento (veja a seção *Estadiamento obrigatório das amostras*, abaixo).
- **Experimentos**
  - Principais campos obrigatórios:
    - `sample_id`: Identificador da Amostra já registrada da qual este experimento foi gerado
    - `analyte_type`: Tipo de analito utilizado para a geração dos dados de sequenciamento
    - `library_kit`: Identificação do kit utilizado para construção de biblioteca
    - `sequencing_center`: Centro onde o sequenciamento foi realizado
    - `sequencing_platform`: Identificação da plataforma utilizada para realizar o sequenciamento das amostras
    - `date_of_sequencing`: Data em que o sequenciamento foi realizado, no formato DD/MM/AAAA
    - `library_layout`: Tipo de leitura utilizada
    - `lane_number`: Número da lane de sequenciamento
    - `fastq_1`: Nome exato do arquivo FASTQ da leitura 1, como consta nos Arquivos de dados do Ensaio
  - Campo condicional: `fastq_2`: Nome exato do arquivo FASTQ da leitura 2. Obrigatório quando `library_layout` = PAIRED.

- Campos utilizados na composição do barcode: `lane_number`. A combinação de `date_of_sequencing`, `sequencing_center` e `sequencing_platform` define o número da corrida (veja **Sistema de identificação e barcode** (cap. 9)).

Estadiamento Obrigatório das Amostras

Cada amostra cadastrada no ATPBR deve informar **pelo menos um sistema de estadiamento**. A plataforma rejeita o cadastro de uma amostra que não tenha **nenhum** dos seguintes campos preenchido:

- **FIGO** (tumores ginecológicos)
- **TNM** (sistema geral)
- **Rai** (leucemia linfocítica crônica)
- **Binet** (leucemia linfocítica crônica)
- **LMC Phase** (leucemia mieloide crônica)
- **AML/ALL Stage** (leucemias agudas)
- **CIN** (neoplasia intraepitelial cervical)

Não é necessário preencher todos: basta o sistema de estadiamento aplicável ao caso. Os campos de estadiamento AJCC são opcionais e não contam para essa regra. Se nenhum dos sistemas acima for informado, a amostra é destacada na tabela de cadastro com a mensagem: *"Informe ao menos um sistema de estadiamento (FIGO, TNM, Rai, Binet, LMC Phase, AML/ALL Stage ou CIN)."*

Datas e tempo: informação em número de dias

Para proteger a identidade dos participantes, o ATPBR **não recebe datas de calendário** de eventos clínicos, como data de nascimento, de diagnóstico, de coleta, de tratamento ou de óbito. Combinadas entre si, essas datas permitiriam reidentificar o paciente.

Em vez disso, os campos temporais recebem o **número de dias** contados a partir de um marco comum, o **diagnóstico**, que corresponde ao **dia 0**. O dia do diagnóstico é a data do primeiro exame que concluiu pela malignidade.

Exemplos:

Campo	O que informar
`diagnosis_age_in_days`	Idade do paciente no diagnóstico, em dias
`days_to_sample`	Dias entre o diagnóstico e a coleta da amostra
`days_to_treatment_first_course`	Dias entre o diagnóstico e o início do primeiro tratamento
`days_to_last_follow_up`	Dias entre o diagnóstico e o último registro do paciente (último seguimento ou óbito)

Por exemplo, se a amostra foi coletada 15 dias depois do diagnóstico, informe `15` em `days_to_sample`.

Para converter as datas do prontuário da sua instituição, use o **conversor de datas em dias**, disponível nos formulários de cadastro da plataforma. A conversão deve ser feita antes do preenchimento, e as datas originais não devem ser enviadas ao ATPBR.

A única data de calendário aceita é a **data de sequenciamento** (`date_of_sequencing`), por não se referir ao paciente. Ela deve estar no formato DD/MM/AAAA, com ano de 4 dígitos, ser uma data válida e não pode estar no futuro.

Versionamento

O dicionário de dados do ATPBR é mantido sob controle de versão e atualizado sempre que necessário para contemplar novas tecnologias, modalidades ômicas, campos clínicos, ontologias de referência e requisitos regulatórios.

Todas as alterações são registradas e documentadas, garantindo a rastreabilidade das mudanças e a compatibilidade com versões anteriores sempre que aplicável.

CAPÍTULO 09

Sistema de identificação e barcode

Vários campos descritos no **Dicionário de dados** (cap. 8), como o acrônimo do projeto, o centro coletor, os códigos de paciente e de amostra e o tipo de ômica, não servem apenas para descrever os dados: eles se combinam para formar o **barcode**, o identificador único de cada paciente, amostra e ensaio. O usuário não digita o barcode; a plataforma o gera automaticamente a partir desses campos.

O ATPBR utiliza um sistema padronizado de identificação para garantir a rastreabilidade de participantes, amostras biológicas, experimentos e dados gerados ao longo de todo o ciclo de vida de um projeto ou estudo. Esse sistema é baseado em barcodes estruturados, que funcionam como identificadores únicos e persistentes dentro da plataforma.

Inspirado no barcode do The Cancer Genome Atlas (TCGA), o sistema de identificação do ATPBR foi adaptado para atender às necessidades específicas da pesquisa oncológica brasileira, incorporando requisitos de governança, desidentificação, integração multiômica e gestão de biobancos.

O barcode constitui o principal mecanismo de ligação entre o material biológico armazenado no Banco Nacional de Tumores (BNT), os metadados registrados na plataforma e os arquivos gerados durante os processos analíticos. Dessa forma, uma mesma identificação acompanha a amostra desde sua coleta até a produção de resultados moleculares e sua disponibilização para pesquisa.

Princípios do Sistema de Identificação

O sistema de barcodes do ATPBR foi desenvolvido com base em quatro princípios fundamentais:

Identificação Única

Cada participante, amostra e ensaio recebe uma combinação única de códigos, eliminando ambiguidades e permitindo a identificação inequívoca dos objetos registrados na plataforma.

Rastreabilidade

Todos os dados produzidos no ATPBR podem ser vinculados à sua origem biológica e experimental. O sistema permite reconstruir o histórico completo de uma amostra, desde sua coleta, armazenamento e processamento até a geração dos arquivos ômicos associados.

Desidentificação

Os identificadores não contêm informações pessoais ou dados que permitam a identificação direta dos participantes, e a correspondência entre os códigos e as pessoas fica guardada apenas na instituição de origem. Como o barcode liga, de propósito, todos os registros de um mesmo participante, os dados são **pseudonimizados**, e não anônimos: por isso, continuam protegidos como dados pessoais sensíveis (veja Governança de dados (cap. 4)).

Interoperabilidade

A estrutura dos barcodes foi projetada para facilitar a integração entre os diferentes módulos do ATPBR e promover compatibilidade com padrões amplamente utilizados em repositórios e plataformas de dados biomédicos.

Material Biológico e Objetos Digitais

O mesmo barcode é utilizado para conectar diferentes componentes da infraestrutura do ATPBR.

No ambiente físico, ele identifica materiais armazenados no Banco Nacional de Tumores, incluindo tecidos, sangue, plasma e outras amostras biológicas.

No ambiente digital, o identificador é utilizado para associar:

- Metadados clínicos.
- Informações experimentais.
- Arquivos FASTQ.
- Arquivos processados.
- Resultados analíticos.
- Registros de processamento e proveniência.

Essa estratégia garante consistência entre as diferentes camadas da plataforma e permite rastrear a origem de qualquer dado disponível no ATPBR.

Estrutura do Barcode

O identificador do ATPBR **não é um código único**: ele é construído de forma **incremental e aninhada**, acompanhando a hierarquia do Modelo de dados (cap. 7) (**Paciente** → **Amostra** → **Ensaio**). Cada nível reaproveita integralmente o barcode do nível anterior e acrescenta os seus próprios campos.

Nível	O que é identificado	Composição	Exemplo
Paciente	o participante do estudo	`PROJETO-CÂNCER-CENTRO-PARTICIPANTE`	`ATPBR-56-001-1234`
Amostra	o material biológico coletado	`‹barcode do paciente›-AMOSTRA-TIPO`	`ATPBR-56-001-1234-5678-01AB`
Ensaio	o experimento ômico sobre a amostra	`‹barcode da amostra›-ENSAIO`	`ATPBR-56-001-1234-5678-01AB-G0101`

Por isso, o barcode "completo" de **sete campos** só existe no nível de **ensaio**: ele é o resultado de empilhar os três níveis. Lido da esquerda para a direita, separado por hífens (`-`), a estrutura completa é:

`PROJETO-CÂNCER-CENTRO-PARTICIPANTE-AMOSTRA-TIPO-ENSAIO`

Exemplo:

`ATPBR-56-001-1234-5678-01AB-G0101`

Cada componente do identificador é descrito nas seções seguintes.

Interpretação do Barcode

Cada segmento do barcode ATPBR possui um significado específico e representa uma informação associada ao projeto, participante, amostra ou ensaio experimental.

Considerando o exemplo:

`ATPBR-56-001-1234-5678-01AB-G0101`

A interpretação é a seguinte:

Campo	Valor	Significado
Projeto	ATPBR	Atlas Tumoral da População Brasileira
Câncer (CID-10)	56	Neoplasia maligna do ovário (C56)
Centro Participante	001	Instituto Nacional de Câncer (INCA)
Participante	1234	Código do participante
Amostra	5678	Identificador da amostra biológica
Tipo de Amostra	01AB	Tecido tumoral maligno congelado
Ensaio	G0101	Sequenciamento de genoma completo (WGS), corrida 01, lane 01

Portanto, o barcode acima identifica uma amostra de tecido tumoral maligno congelado proveniente de uma paciente com câncer de ovário, coletada no INCA no contexto do projeto ATPBR e submetida a um ensaio de sequenciamento de genoma completo.

Significado dos Códigos

- **Código do Projeto**

É o **acrônimo do projeto**, definido pelo usuário no cadastro do Projeto. Deve ter de **5 a 7 letras maiúsculas** e é único na plataforma: cada projeto tem o seu.

Exemplo:

Código	Projeto
ATPBR	Atlas Tumoral da População Brasileira

- **Código do Câncer (CID-10)**

Identifica o diagnóstico oncológico do participante pela **categoria da Classificação Internacional de Doenças (CID-10)**, representada pelos 2 dígitos do código, sem a letra "C". Por exemplo, a categoria C56 (neoplasia maligna do ovário) entra no barcode como `56`.

> **Como esse código é definido:** o usuário **não digita** o código. No **cadastro do Paciente**, ele informa o tipo de câncer (classificação OncoTree, entre os tipos do tecido definido no Estudo), o subtipo, quando conhecido, e o órgão de origem do tumor. A partir dessas respostas, a plataforma calcula automaticamente a categoria CID-10 correspondente, e é esse valor que entra no barcode. Nas neoplasias hematológicas (leucemias, linfomas, mieloma), a categoria é definida pelo subtipo, independentemente do órgão.

Exemplos de códigos:

Código	Categoria CID-10	Descrição
15	C15	Esôfago
16	C16	Estômago
18	C18	Cólon
20	C20	Reto
34	C34	Brônquios e pulmão
43	C43	Melanoma
44	C44	Outras neoplasias de pele
50	C50	Mama
53	C53	Colo do útero
56	C56	Ovário
61	C61	Próstata
91	C91	Leucemia linfoide
92	C92	Leucemia mieloide
99	(não se aplica)	Neoplasia não maligna (benigna, in situ ou de comportamento incerto)

São aceitas as categorias de neoplasias malignas da CID-10 (C00 a C97), além do código `99` para neoplasias não malignas. A lista completa está disponível na planilha de metadados. Os códigos são estáveis: os já em uso nunca são alterados.

- **Código do Centro Participante**

Identifica o **centro coletor**, a instituição responsável pela coleta ou submissão da amostra. Cada centro recebe um código de **3 dígitos**.

Como o centro entra no sistema:

1. A instituição é **cadastrada pelo usuário na criação do Projeto**, junto com o seu identificador **ROR**.
2. A partir daí, ela passa a aparecer no **vocabulário controlado "Collector Center"** e fica disponível para seleção **no cadastro de um paciente**.
3. O código gravado no barcode é o da instituição selecionada nesse vocabulário.

Os códigos são globais e permanentes:

- Cada instituição tem **um único código fixo em toda a plataforma**, identificada de forma única pelo seu **ROR id**, de modo que o mesmo centro real sempre resolve para o mesmo código.
- Um **novo código só é gerado quando a instituição ainda não foi selecionada na plataforma antes**; centros já utilizados reaproveitam o código existente.
- Os códigos **nunca são reutilizados nem removidos**, para que os barcodes de amostras já cadastradas permaneçam válidos.

Centros pré-definidos:

Código	Instituição	ROR id
001	Instituto Nacional de Câncer (INCA)	055n68305
002	Instituto do Câncer do Estado de São Paulo (ICESP)	005vqqr19
003	Instituto do Câncer do Ceará (ICC)	01pg9bv14

Novos centros recebem o próximo código sequencial disponível, na ordem em que são utilizados pela primeira vez.

- **Código do Participante**

Código de **4 dígitos** que identifica o participante. É **definido pelo centro coletor** e informado no campo `patient_id` do cadastro do paciente.

Exemplo:

Código	Significado
1234	Código do participante

O código não deve ter qualquer relação com dados pessoais do participante: não use, por exemplo, o número do prontuário nem partes da data de nascimento.

- **Código da Amostra**

Código de **4 dígitos** que identifica a amostra biológica vinculada a um participante. Também é **definido pelo centro coletor** e informado no campo `sample_id` do cadastro da amostra.

Exemplo:

Código	Significado
5678	Amostra biológica específica

Um mesmo participante pode possuir múltiplas amostras. A plataforma não aceita dois registros com o mesmo barcode e avisa quando detecta um possível cadastro em duplicidade.

- **Código do Tipo de Amostra**

O campo Tipo de Amostra é composto por quatro caracteres e descreve simultaneamente o material biológico, o status patológico e a forma de preservação da amostra.

Estrutura:

`TTSP`

Onde:

- **TT** = Tipo de material (2 dígitos)
- **S** = Status da amostra (1 letra)
- **P** = Preservação da amostra (1 letra)

Por exemplo, `01AB` corresponde a:

- **01** = Tecido (*Tissue*)
- **A** = Tumoral maligno (*Malignant Tumor*)
- **B** = Congelado (*Frozen*)

Resultado: **Tecido tumoral maligno congelado**.

> **Idioma dos valores:** na plataforma, os rótulos destes vocabulários aparecem **em inglês**, para manter a conformidade entre todos os projetos. As tabelas abaixo trazem o rótulo oficial (EN) e a tradução em português (PT). O que entra no barcode é sempre o **código** (a coluna da esquerda), não o texto.

Códigos de Tipo de Material

Código	Sample Type (EN)	Tipo de Material (PT)
01	Tissue	Tecido
02	Blood	Sangue
03	Plasma	Plasma
04	Serum	Soro
05	Bone Marrow	Medula Óssea
06	Peritoneal Fluid	Líquido Peritoneal
07	Synovial Fluid	Líquido Sinovial
08	Pericardial Fluid	Líquido Pericárdico
09	Cerebrospinal Fluid	Líquido Cefalorraquidiano
10	Bronchial Lavage	Lavado Brônquico
12	Urine	Urina

Código	Sample Type (EN)	Tipo de Material (PT)
13	Saliva	Saliva
14	Sputum	Escarro
15	Feces	Fezes
16	Skin	Pele
17	Cells	Células
18	Breast Secretion	Secreção Mamária
19	Pleural Fluid	Líquido Pleural
20	Extracellular Vesicles	Vesículas Extracelulares / Exossomos
21	Aspiration Biopsy	Biópsia Aspirativa
22	Buffy Coat	Camada Leucoplaquetária (Buffy Coat)

O código 11 não é utilizado.

Códigos de Status da Amostra

Código	Sample Status (EN)	Status da Amostra (PT)
A	Malignant Tumor	Tumoral Maligno
B	Benign Tumor	Tumoral Benigno
C	Lesion	Lesão
D	Metastasis	Metástase
E	Normal - Adjacent	Normal Adjacente
F	Normal - Distant	Normal Distante
G	Lymph Node - Metastatic	Linfonodo Metastático
H	Lymph Node - Reactive	Linfonodo Reacional
I	Inflammation	Inflamação
J	Cell Line	Linhagem Celular
K	Normal Non Tumoral	Normal Não Tumoral

Códigos de Preservação da Amostra

Código	Sample Preservation (EN)	Preservação da Amostra (PT)
A	Fresh	Fresco
B	Frozen	Congelado

Código	Sample Preservation (EN)	Preservação da Amostra (PT)
C	FFPE (Formalin-Fixed, Paraffin-Embedded)	FFPE (fixado em formalina e incluído em parafina)
D	Paraformaldehyde (PFA)	Paraformaldeído (PFA)
E	RNAlater	RNAlater
F	Culture Medium	Meio de Cultura
G	Saline Buffer	Tampão Salino
L	Saline Solution	Solução Salina

> **Anticoagulantes:** o aditivo do tubo de coleta (por exemplo, EDTA ou citrato) é informado em um campo próprio, `collection_tube_additive`, que é opcional e **não entra no barcode**.

Exemplos de Interpretação

Código	Interpretação
01AB	Tecido tumoral maligno congelado
01AC	Tecido tumoral maligno preservado em FFPE
01EB	Tecido normal adjacente congelado
02KB	Sangue normal não tumoral congelado
22KB	Buffy coat normal não tumoral congelado
17JF	Linhagem celular mantida em meio de cultura
20AB	Vesículas extracelulares tumorais malignas congeladas

A combinação desses códigos permite descrever de forma compacta e padronizada as principais características biológicas e operacionais de cada amostra, mantendo a rastreabilidade ao longo de todo o fluxo de trabalho do ATPBR.

- **Código do Ensaio**

O campo **Ensaio** identifica a modalidade ômica utilizada e informações associadas à execução experimental. Esse campo permite distinguir diferentes experimentos realizados sobre uma mesma amostra e manter a rastreabilidade dos dados gerados.

Estrutura:

`OCCLL`

Onde:

- **O** = Modalidade ômica (1 caractere)
- **CC** = Número da corrida experimental (2 dígitos)
- **LL** = Número da lane (2 dígitos)

Por exemplo:

`G0101`

corresponde a:

- **G** = Whole Genome Sequencing (WGS)
- **01** = Primeira corrida
- **01** = Lane 1

Resultado:

Sequenciamento de Genoma Completo (WGS), corrida 01, lane 01

Códigos de Modalidades Ômicas

Código	Modalidade Ômica (EN)	Tradução (PT)
G	Whole Genome Sequencing (WGS)	Sequenciamento de Genoma Completo
X	Whole Exome Sequencing (WES)	Sequenciamento de Exoma Completo
R	RNA-Seq	Sequenciamento de RNA
P	Gene Panel Sequencing	Sequenciamento por Painel Gênico

Os códigos M (metilação de DNA) e B (metilação de DNA 5-base) estão reservados para versões futuras (veja as modalidades aceitas em Tipos de dados (cap. 10)).

Número da Corrida

Número sequencial **atribuído automaticamente pela plataforma**. O usuário **não define** esse número manualmente.

A corrida é identificada pela combinação de três campos informados em cada Experimento: **data de sequenciamento**, **centro de sequenciamento** e **plataforma de sequenciamento**. A numeração é feita dentro de cada **Estudo**:

- A primeira combinação registrada no Estudo recebe `01`, a segunda recebe `02`, e assim por diante.
- Experimentos com a mesma data, o mesmo centro e a mesma plataforma pertencem à mesma corrida e recebem o mesmo número.
- Uma vez atribuído, o número da corrida não muda.

Por isso, esses três campos são obrigatórios e devem ser preenchidos com cuidado. Além de distinguir repetições técnicas e novas execuções do mesmo experimento, essa identificação **facilita análises de efeito de lote (batch) posteriores**, em que é importante saber quais dados foram gerados em corridas diferentes.

Código	Significado
01	Primeira corrida
02	Segunda corrida
03	Terceira corrida

Código	Significado
...	Corridas subsequentes

Número da Lane

Representa a lane de sequenciamento utilizada durante a corrida.

Código	Significado
01	Lane 1
02	Lane 2
03	Lane 3
04	Lane 4
...	Demais lanes

A identificação da lane auxilia na rastreabilidade técnica e na investigação de possíveis problemas relacionados ao sequenciamento.

Exemplos de Interpretação

Código	Interpretação
G0101	WGS, corrida 01, lane 01
X0102	WES, corrida 01, lane 02
R0201	RNA-Seq, corrida 02, lane 01
P0101	Painel gênico, corrida 01, lane 01

Exemplos de Barcodes Completos

Barcode	Interpretação
ATPBR-56-001-1234-5678-01AB-G0101	Amostra de tecido tumoral maligno congelado de câncer de ovário submetida a WGS na corrida 01, lane 01
ATPBR-50-002-3456-7890-01AC-R0101	Amostra de tecido tumoral maligno preservado em FFPE de câncer de mama submetida a RNA-Seq na corrida 01, lane 01
ATPBR-61-003-4321-8765-02KB-X0201	Amostra de sangue normal não tumoral congelado de câncer de próstata submetida a WES na corrida 02, lane 01

A combinação do código de ensaio com os demais campos do barcode permite identificar de forma única o contexto biológico e experimental associado aos dados produzidos no ATPBR, garantindo rastreabilidade desde a amostra física até os arquivos gerados pelo sequenciamento.

Referências

- National Cancer Institute. Genomic Data Commons. *TCGA Barcode*. GDC Docs. Disponível em: https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

CAPÍTULO 10

Tipos de dados

Os capítulos anteriores trataram dos **metadados**: como se organizam (Modelo de dados (cap. 7)), como se preenchem (Dicionário de dados (cap. 8)) e como geram os identificadores (Sistema de identificação e barcode (cap. 9)). Este capítulo trata dos **formatos**: o que o ATPBR aceita na submissão e o que a plataforma gera a partir disso.

Visão Geral

O ATPBR aceita dois tipos de dados para submissão:

Categoria	Formato aceito
Metadados	Formulários do portal ou planilhas-modelo Excel (.xlsx ou .xls)
Dados brutos de sequenciamento	FASTQ compactado com gzip (.fastq.gz ou .fq.gz)

Todos os demais arquivos são gerados internamente pela plataforma e **não devem ser submetidos**.

Metadados

Os metadados podem ser registrados de duas formas:

- **Formulários do portal:** cadastro de um registro por vez (um paciente, uma amostra ou um experimento). Projetos, Investigações, Estudos e Ensaios são sempre cadastrados por formulário.
- **Planilhas-modelo:** cadastro de vários pacientes, amostras ou experimentos de uma só vez. Baixe o modelo na própria página de cadastro, preencha e envie o arquivo. Antes de salvar, uma prévia mostra as linhas lidas e os erros encontrados.

Ao usar planilhas, utilize exclusivamente os templates disponibilizados pelo ATPBR e mantenha sua estrutura original (não renomeie, remova ou reordene colunas). As regras de preenchimento de cada campo são as mesmas dos formulários e estão no Dicionário de dados (cap. 8).

Dados Brutos de Sequenciamento

Os dados de sequenciamento devem corresponder aos arquivos primários gerados pelos equipamentos de sequenciamento.

Modalidades aceitas

Nesta primeira versão, são aceitas as seguintes modalidades ômicas:

- Sequenciamento de genoma completo (**WGS**).
- Sequenciamento de exoma completo (**WES**).
- Sequenciamento por **painel gênico**.
- Sequenciamento de RNA (**RNA-seq**).

Outras modalidades, como metilação de DNA, serão incorporadas em versões futuras.

Requisitos dos arquivos FASTQ

- Arquivos compactados no formato gzip (.fastq.gz ou .fq.gz).
- Em leituras pareadas (paired-end), os dois arquivos de cada amostra: leitura 1 e leitura 2.
- Vinculados a um Ensaio: enviados pelo portal ou localizados na Plataforma Multiusuário de Bioinformática do INCA (veja **Submissão de dados** (cap. 5)).
- Nomes idênticos aos informados nos campos `fastq_1` e `fastq_2` dos Experimentos.
- Nomes de arquivo sem dados pessoais, como nome ou número de prontuário do paciente.
- Integridade verificada por checksum.

Dados Gerados pelo ATPBR

A partir dos FASTQ, a plataforma gera internamente:

- Arquivos de alinhamento (CRAM).
- Variantes somáticas (VCF e MAF).
- Matrizes de expressão gênica.
- Dados de número de cópias.
- Métricas de controle de qualidade.

Os pipelines que geram esses arquivos estão descritos no próximo capítulo.

PARTE 4

Análise e uso dos dados

O processamento dos dados, a exploração dos resultados e as responsabilidades de quem os utiliza.

11 Plataforma analítica

12 Navegador de resultados

13 Políticas de uso

CAPÍTULO 11

Plataforma analítica

Os arquivos FASTQ aceitos na submissão (veja **Tipos de dados** (cap. 10)) são processados pela **Plataforma Analítica do ATPBR (ATPBR-AP)**. Este capítulo descreve quais pipelines são usados, como estão configurados e quais resultados geram.

Visão Geral

A execução dos pipelines é **gerenciada centralmente pela equipe do ATPBR**: o usuário não precisa configurar nem executar nenhuma análise. Isso garante que todos os dados sejam processados da mesma forma, o que permite comparar estudos diferentes.



Os pipelines são fluxos de trabalho da comunidade nf-core, construídos em Nextflow, executados em contêineres e com versões e parâmetros fixos.

Dados	Pipeline	Principal resultado
WGS, WES e painel gênico	`nf-core/sarek` 3.10.0	Variantes somáticas anotadas e filtradas
RNA-seq	`nf-core/rnaseq` 3.18.0	Matrizes de expressão gênica

Controle de Qualidade

O controle de qualidade permite avaliar as características das leituras de sequenciamento e identificar possíveis problemas que possam comprometer as análises subsequentes.

As ferramentas utilizadas incluem:

- FastQC: avalia a qualidade das leituras brutas, incluindo qualidade por base, conteúdo GC, duplicação e presença de adaptadores.
- fastp: realiza filtragem e pré-processamento das leituras, incluindo remoção de adaptadores e sequências de baixa qualidade, conforme os parâmetros configurados.
- MultiQC: consolida métricas e relatórios de diferentes ferramentas, facilitando a avaliação conjunta das amostras.

Os resultados das duas ferramentas são reunidos em um relatório **MultiQC**. As principais métricas avaliadas são:

Métrica	O que se espera
Qualidade por base (Phred)	A maior parte das bases na zona verde (Q ≥ 30). Uma pequena queda nas primeiras e nas últimas bases da leitura é normal.
Conteúdo GC	Distribuição em forma de sino, próxima de 41% no DNA e de 50% no mRNA (Piovesan et al., 2019; Palma et al., 2025). Vários picos sugerem contaminação.
Bases não identificadas (N)	Menos de 5%. Valores maiores indicam problema na qualidade do sequenciamento.
Duplicação	A interpretação depende do tipo de ensaio, profundidade e complexidade da biblioteca. Níveis elevados podem ser esperados em WES, painéis gênicos e RNA-seq.
Conteúdo de adaptadores	Menos de 0,1% após o corte pelo fastp.
Leituras filtradas (fastp)	O fastp descarta leituras com mais de 40% das bases com Q < 15, com mais de 5 bases N, com menos de 15 pb após o corte ou formadas por sequências repetitivas (por exemplo, caudas poli-G).

Análise de WGS, WES e Painel Gênico

O processamento de dados de sequenciamento de DNA é realizado utilizando o `nf-core/sarek` , versão 3.10.0, com configurações específicas do ATPBR. O fluxo compreende as seguintes etapas:

1. **Controle de qualidade e pré-processamento** com FastQC e fastp.
2. **Alinhamento** ao genoma de referência GRCh38 com BWA-MEM. As leituras de diferentes lanes são reunidas por amostra e salvas em formato CRAM.
3. **Pré-processamento** dos alinhamentos com o GATK: marcação de duplicatas e recalibração da qualidade das bases.
4. **Chamada de variantes somáticas**identificação de variantes de nucleotídeo único (SNVs) e pequenas inserções e deleções (indels) utilizando o Mutect2, considerando duas estratégias:
    - **Amostras pareadas (tumor-normal):** utiliza dados do tumor e de uma amostra normal correspondente ao mesmo paciente, permitindo distinguir variantes somáticas de variantes germinativas com maior precisão.
    - **Amostras não pareadas (tumor-only):** utiliza exclusivamente os dados tumorais, sem uma amostra normal correspondente. Nesse caso, a identificação de variantes somáticas depende mais fortemente de filtros populacionais e de recursos de identificação de artefatos, podendo permanecer variantes germinativas entre os resultados.
5. **Anotação** dos efeitos de cada variante em genes, transcritos e proteínas com o Ensembl VEP.
6. **Filtragem** de artefatos e de variantes potencialmente germinativas, considerando a estratégia de análise:
    - **Amostras pareadas:** a comparação tumor-normal permite identificar e excluir variantes presentes na amostra normal correspondente, complementada pelos filtros de qualidade do Mutect2.

- **Amostras não pareadas:** são filtradas variantes presentes no painel de normais (PoN) do 1000 Genomes, variantes com frequência populacional acima de 1% no gnomAD e variantes com frequência acima de 0,1% na população brasileira, segundo o ABraOM. A filtragem pelo ABraOM é realizada por um pipeline interno da equipe do ATPBR, após o `nf-core/sarek`.

Configuração

`nf-core/sarek`	Versão 3.10.0
Genoma de referência	*Homo sapiens* iGenomes GATK GRCh38
Corte (trimming)	fastp
Alinhador	BWA-MEM
Chamador de variantes	Mutect2 (tumor-normal e tumor-only)
Anotação funcional	Ensembl VEP
Filtros	PoN, gnomAD, 1000 Genomes e ABraOM, conforme a etapa e a estratégia

Pós-processamento das variantes somáticas

Após a execução do `nf-core/sarek`, scripts desenvolvidos pela equipe do ATPBR processam os arquivos VCF anotados para selecionar variantes, organizar informações genômicas e incorporar dados populacionais complementares.

O processamento documentado compreende as seguintes etapas:

1. **Seleção das variantes PASS:** São selecionadas as variantes cujo campo FILTER no VCF apresenta o valor PASS, indicando que atenderam aos critérios de filtragem aplicados anteriormente.
2. **Conversão para tabela:** os campos da anotação do VEP e as contagens de leituras do tumor (profundidade DP, leituras por alelo AD e frequência alélica AF) são extraídos para uma tabela por amostra. A frequência do 1000 Genomes informada pelo VEP passa a se chamar `GMAF`.
3. **Anotação com o ABraOM:** cada variante recebe a frequência alélica observada na população brasileira.

Análise de RNA-seq

O pipeline `nf-core/rnaseq` processa RNA-seq de leituras curtas (RNA total, enriquecido com poli(A) ou depletado de rRNA) nas seguintes etapas:

1. **Concatenação** dos FASTQ de uma mesma amostra sequenciada em várias lanes.
2. **Controle de qualidade e corte** com FastQC e fastp.
3. **Pseudoalinhamento e quantificação** com o Salmon, que atribui as leituras diretamente aos transcritos conhecidos do transcriptoma de referência. A expressão é estimada por transcrito e somada por gene.
4. **Normalização** em Transcritos por Milhão (TPM), que corrige as diferenças de tamanho de biblioteca e de comprimento dos transcritos.

Por padrão, não são aplicadas etapas adicionais de filtragem. Se o controle de qualidade indicar problemas, a equipe pode ajustar parâmetros específicos (veja *Limitações*, abaixo).

Configuração

`nf-core/rnaseq`	**Versão 3.18.0**
Genoma de referência	*Homo sapiens* Ensembl Release 113 GRCh38
Corte (trimming)	fastp
Pseudoalinhamento e quantificação	Salmon

Limitações

- **Modalidades:** apenas as modalidades listadas em **Tipos de dados** (cap. 10) são processadas.
- **Organismo:** os pipelines são configurados apenas para dados humanos. Amostras de organismos não humanos ou mistas, como xenotransplantes derivados de pacientes, não são suportadas.
- **Variantes germinativas:** não há análise exclusivamente germinativa, pois o Mutect2 é otimizado para variantes somáticas. Além disso, variantes germinativas são dados de acesso controlado (veja **Acesso aos dados** (cap. 6)).
- **Heterogeneidade dos dados:** os dados vêm de vários centros, com diferenças de manuseio das amostras, kits de biblioteca e protocolos de sequenciamento. Para incluir amostras raras com qualidade comprometida, a equipe pode ajustar os parâmetros do pipeline. Esses ajustes, assim como diferenças entre corridas, podem introduzir **efeitos de lote**. Por isso, o número da corrida faz parte do barcode (veja **Sistema de identificação e barcode** (cap. 9)).

CAPÍTULO 12

Navegador de resultados

cBioPortal do ATPBR

Os resultados gerados pela Plataforma analítica (cap. 11) chegam aos usuários pelo **cBioPortal do ATPBR**, um ambiente interativo para navegar, visualizar e analisar dados genômicos e clínicos de câncer sem exigir conhecimento avançado em bioinformática. Este capítulo descreve o que está disponível no portal e como explorá-lo.

O ATPBR utiliza o **cBioPortal for Cancer Genomics, versão 7.2**, uma plataforma de código aberto amplamente utilizada pela comunidade científica internacional para exploração integrada de dados moleculares e clínicos de câncer.

O portal é voltado a pesquisadores translacionais, biólogos moleculares, clínicos, bioinformatas e cientistas de dados interessados no panorama molecular dos tumores brasileiros.

1. Dados disponíveis

O cBioPortal integra, em um mesmo estudo, as camadas de dados geradas ao longo do fluxo do ATPBR:

Camada	Origem	O que permite explorar
Mutações somáticas	Variantes filtradas e anotadas pelo `nf-core/sarek` (WGS, WES, painel gênico)	Genes alterados de forma recorrente, tipos de mutação e distribuição das alterações entre os casos.
Número de cópias	Dados segmentados de número de cópias, quando disponíveis	Ganhos e perdas genômicas, amplas ou focais.
Expressão gênica	Matrizes em TPM do `nf-core/rnaseq` (RNA-seq)	Diferenças de expressão entre genes, amostras e grupos, e efeito transcricional de alterações genômicas.
Dados clínicos e de amostras	Metadados cadastrados na submissão (veja Dicionário de dados (cap. 8))	Filtragem de coortes, estratificação das análises e interpretação clínica dos achados.

Nem todo estudo terá todas essas camadas: a disponibilidade depende dos ensaios realizados, da qualidade das amostras e da etapa de disponibilização. Consulte a descrição de cada estudo no portal.

2. Como navegar

O caminho típico de exploração tem quatro passos:

1. **Escolher um estudo** na lista de estudos. Cada estudo corresponde a um conjunto de dados curado, geralmente um tipo de tumor.
2. **Conhecer a coorte na Study View**, que resume o número de pacientes e amostras, os tipos de dados disponíveis e a distribuição das principais variáveis clínicas e moleculares.
3. **Filtrar a coorte** pelos atributos disponíveis, como tipo de tumor, histologia, tipo de amostra, estadiamento ou disponibilidade de sequenciamento.
4. **Consultar genes** de interesse e examinar os resultados nas visualizações descritas abaixo.

Os estudos e perfis visíveis variam conforme as permissões de cada usuário e a etapa de disponibilização dos dados (veja Acesso aos dados (cap. 6)).

3. Principais funcionalidades

Funcionalidade	Para que serve
Study View	Visão geral do estudo e ponto de partida para aplicar filtros na coorte.
Consulta por gene	Buscar um ou mais genes na coorte selecionada e ver como estão alterados. Por padrão, inclui mutações não sinônimas, fusões, amplificações e deleções profundas.
OncoPrint	Resumo visual, amostra por amostra, das alterações nos genes consultados. Ajuda a identificar eventos recorrentes, alterações concomitantes e mutuamente exclusivas.
Visualização de mutações	Detalhes de cada variante: tipo, posição na proteína e amostras afetadas. Útil para identificar hotspots.
Número de cópias e expressão	Explorar amplificações, deleções e níveis de expressão, quando esses dados existem no estudo.
Aba Plots	Comparar medidas moleculares com anotações clínicas, por exemplo a expressão de um gene por subtipo tumoral.
Comparação de grupos	Comparar características moleculares e clínicas entre subcoortes definidas pelo usuário.
Análise de sobrevida	Gráficos de desfecho, quando o estudo inclui status vital e tempo de seguimento.
Visão do paciente / da amostra	Todas as informações clínicas e moleculares de um caso, incluindo várias amostras do mesmo paciente.
Onco Query Language (OQL)	Consultas mais precisas, especificando classes de mutação, variantes exatas, estados de número de cópias ou limiares de expressão.
Download e compartilhamento	Exportar tabelas e resultados para análise em R ou Python, e salvar ou compartilhar consultas. Depende das permissões do usuário.

4. Casos de uso

O portal é especialmente útil para:

- Explorar a frequência e a distribuição de alterações somáticas em genes selecionados nas coortes brasileiras.
- Comparar padrões moleculares entre subgrupos clínicos, como subtipo, estadiamento ou histologia.
- Investigar alterações concomitantes ou mutuamente exclusivas que sugiram vias biológicas compartilhadas.
- Gerar hipóteses sobre biomarcadores e assinaturas moleculares para validação futura.
- Verificar, antes de uma análise externa, se uma coorte tem os perfis moleculares necessários.
- Revisar casos individuais em seu contexto molecular e clínico.
- Preparar figuras e resumos para reuniões, relatórios e o planejamento de publicações.

5. Como os dados chegam ao portal

Os estudos são preparados e carregados exclusivamente pela equipe de dados do ATPBR. Os resultados dos pipelines e os metadados cadastrados são harmonizados, convertidos para os formatos exigidos pelo cBioPortal e validados antes do carregamento. Quando novos ensaios ou correções ficam disponíveis, o estudo é revisado, revalidado e recarregado.

Cada estudo no portal é formado pelos seguintes grupos de arquivos:

Grupo de arquivos	Função no portal
Metadados do estudo	Identificador, nome, descrição, citação e visibilidade do estudo.
Dados clínicos do paciente	Atributos no nível do paciente (Tabela 1).
Dados clínicos da amostra	Atributos no nível da amostra (Tabela 2).
Perfil de mutações	Mutações somáticas.
Perfil de expressão	Expressão de mRNA e de transcritos.
Número de cópias	Dados segmentados de número de cópias.
Listas de casos	Amostras disponíveis para cada tipo de análise.

Tabela 1. Campos no nível do paciente.

Atributo	Tipo	Descrição
PATIENT_ID	Texto	Identificador do paciente, que vincula seus dados clínicos às amostras e aos dados moleculares.
SAMPLES_PER_PATIENT	Número	Número de amostras do paciente no estudo.
SEX	Texto	Sexo do paciente.
RACE	Texto	Raça/cor do paciente, em categorias agrupadas.
COUNTRY	Texto	País associado ao paciente, conforme a definição do estudo.
DIAGNOSIS_AGE	Número	Idade ao diagnóstico, em faixas de 5 anos.

Atributo	Tipo	Descrição
OS_STATUS	Texto	Status de sobrevida global (vivo ou falecido).
OS_MONTHS	Número	Tempo de sobrevida global em meses arredondados, a partir do diagnóstico.

Tabela 2. Campos no nível da amostra.

Atributo	Tipo	Descrição
PATIENT_ID	Texto	Paciente ao qual a amostra pertence.
SAMPLE_ID	Texto	Identificador da amostra, que a vincula aos dados moleculares.
CANCER_TYPE	Texto	Tipo de câncer.
CANCER_TYPE_DETAILED	Texto	Tipo de câncer detalhado.
ONCOTREE_CODE	Texto	Código OncoTree da doença.
SAMPLE_TYPE	Texto	Tipo de espécime, como tumor primário ou metástase.
TUMOR_STAGE_2009	Texto	Estadiamento do tumor.
TREATMENT_DRUG_NAME	Texto	Medicamento ou esquema de tratamento, quando disponível.
TREATMENT_RESPONSE	Texto	Resposta ao tratamento.
WGS_DATA	Texto	Indica se a amostra tem dados de sequenciamento do genoma completo (WGS).
WES_DATA	Texto	Indica se a amostra tem dados de sequenciamento do exoma (WES).
PANEL_DATA	Texto	Indica se a amostra tem dados de painel genético.
RNASEQ_DATA	Texto	Indica se a amostra tem dados de RNA-seq.
METHYLATION_DATA	Texto	Indica se a amostra tem dados de metiloma.

6. Privacidade

Todos os dados do portal são desidentificados antes da disponibilização, conforme os princípios apresentados em **Sobre o ATPBR** (cap. 1). Desidentificados não quer dizer anônimos: por isso, só são publicados os estudos validados e aprovados para divulgação, depois de uma avaliação dos riscos de reidentificação (veja **Governança de dados** (cap. 4)). O uso dos dados obtidos no portal, incluindo a proibição de tentativas de reidentificação e as regras de citação, segue as **Políticas de uso** (cap. 13). Os resultados devem ser interpretados considerando o escopo e as limitações de cada coorte.

CAPÍTULO 13

Políticas de uso

O acesso aos dados do ATPBR, seja aberto pelo **Navegador de resultados** (cap. 12) ou controlado (veja **Acesso aos dados** (cap. 6)), está sujeito às políticas descritas a seguir. Elas complementam os princípios de desidentificação e de conformidade com a LGPD apresentados em **Sobre o ATPBR** (cap. 1) e as regras de **Governança de dados** (cap. 4).

Responsabilidades dos usuários

Os usuários autorizados comprometem-se a:

- Utilizar os dados exclusivamente na pesquisa, no prazo e nas condições aprovadas;
- Não tentar identificar participantes, seus familiares ou instituições de origem, nem combinar os dados do ATPBR com outras fontes para esse fim;
- Não contatar participantes a partir dos registros recebidos;
- Não compartilhar credenciais, não disponibilizar cópias a pessoas não autorizadas e não retransmitir os dados a terceiros;
- Não utilizar os dados para finalidades não autorizadas, inclusive assistenciais ou diagnósticas. O uso comercial só é admitido quando o consentimento dos participantes o permite (conjuntos sem os códigos DUO NCU ou NPU), nas mesmas regras de acesso, sem exclusividade nem patente sobre os dados e com compartilhamento de benefícios, e os casos excepcionais são decididos pelo CAD com o CGN;
- Não utilizar os dados para identificação forense ou policial, para seleção ou precificação de risco por seguradoras ou empregadores, para publicidade dirigida, nem para sustentar hierarquias raciais, étnicas ou de ancestralidade;
- Não remover controles de segurança nem transferir os dados para ambientes não autorizados;
- Adotar medidas de segurança proporcionais aos riscos, como autenticação individual, limitação de privilégios, proteção das credenciais, uso dos dados exclusivamente na instância segura do ATPBR e nenhuma tentativa de extrair dados individuais;
- Respeitar as restrições de origem, que continuam valendo para os dados derivados, e eventuais moratórias de publicação;
- Não publicar resultados que permitam identificar, direta ou indiretamente, os participantes. A autorização para analisar os dados não equivale a autorização para publicar registros individuais;
- Citar o ATPBR e os estudos de origem, reconhecendo os centros que contribuíram com amostras e dados, em publicações e produtos resultantes da utilização dos dados (o artigo de referência está em

Sobre o ATPBR (cap. 1)).

Os usuários de dados de acesso controlado também se comprometem a:

- Comunicar ao INCA, pelo canal e no prazo previstos no instrumento de uso, incidentes de segurança, suspeitas de reidentificação ou de uso indevido;
- Comunicar previamente mudanças relevantes na finalidade, nos métodos, na equipe, nas instituições participantes, no ambiente de tratamento ou nos conjuntos solicitados, que podem exigir nova decisão do Comitê de Acesso a Dados;
- Ao fim do prazo ou na revogação do acesso, eliminar ou devolver os dados, ou mantê-los apenas na forma autorizada, e confirmar essa providência ao ATPBR.

O descumprimento dessas condições poderá resultar na suspensão ou revogação do acesso e em outras medidas previstas no instrumento de uso de dados e nas políticas do ATPBR. A aplicação dessas medidas cabe ao Comitê de Acesso a Dados (CAD), e as decisões ficam registradas.

O acesso a dados não inclui o acesso às amostras biológicas correspondentes, cuja governança é de competência do Banco Nacional de Tumores (BNT) e não é regulada pelo ATPBR.

Transparência e rastreabilidade

As solicitações, decisões, anuências, habilitações, acessos, suspensões, revogações e encerramentos são registrados, com proteção contra alteração indevida, para fins de auditoria, rastreabilidade e conformidade com as políticas de governança do ATPBR (veja **Governança de dados** (cap. 4)). Uma lista pública dos pedidos aprovados, com o título do projeto, a instituição e os conjuntos autorizados, é divulgada periodicamente no site do ATPBR.

CAPÍTULO 14

Perguntas frequentes

Esta página reúne as dúvidas mais comuns sobre o uso do ATPBR. Se a sua pergunta não estiver aqui, entre em contato com a equipe do ATPBR pelo e-mail atpbr@inca.gov.br.

Sobre o ATPBR

? O ATPBR recebe dados que identificam os pacientes?

Não diretamente. O ATPBR recebe apenas dados previamente desidentificados pelas instituições participantes: nomes, documentos, números de prontuário, contatos e datas de eventos clínicos não devem ser enviados, e a correspondência entre os códigos e as pessoas fica somente na instituição de origem. Mesmo assim, os dados são pseudonimizados, e não anônimos, e por isso são protegidos como dados pessoais sensíveis. Saiba mais em **Governança de dados** (cap. 4) e **Políticas de uso** (cap. 13).

? Quais tipos de dados ômicos são aceitos?

WGS, WES, painel gênico e RNA-seq. Outras modalidades, como metilação de DNA, serão incorporadas em versões futuras. Veja **Tipos de dados** (cap. 10).

Cadastro e submissão

? Em que ordem devo cadastrar as informações?

Projeto → Investigação → Estudo → Pacientes → Amostras → Ensaio → Arquivos de dados → Experimentos. Cada etapa depende da anterior. O passo a passo está em **Submissão de dados** (cap. 5), e o motivo dessa ordem, em **Modelo de dados** (cap. 7).

? Posso cadastrar vários pacientes, amostras ou experimentos de uma vez?

Sim, por meio das planilhas-modelo (Excel, `.xlsx` ou `.xls`), que podem ser baixadas na própria página de cadastro. Veja **Tipos de dados** (cap. 10).

? Meu centro coletor não aparece na lista do cadastro de pacientes. O que faço?

O centro coletor precisa estar entre as **Instituições Associadas** do Projeto. Edite o Projeto e inclua a instituição; a partir daí, ela passa a aparecer na lista "Collector Center" do cadastro de pacientes. Veja **Sistema de identificação e barcode** (cap. 9).

Por que não consigo criar um segundo Estudo com o mesmo tipo de câncer?

Dentro de um mesmo projeto, cada tipo de câncer só pode ter um Estudo. Novos pacientes com esse tipo de câncer devem ser cadastrados no Estudo já existente.

Por que o tipo de câncer que procuro não aparece no cadastro do paciente?

As opções de tipo de câncer do paciente se limitam ao tecido definido no Estudo. Se o tipo procurado pertence a outro tecido, ele deve ser cadastrado no Estudo correspondente.

Datas e tempo

Por que não posso informar datas, como a data do diagnóstico ou da coleta?

Combinadas entre si, as datas de eventos clínicos permitem reidentificar o paciente. Por isso, os campos temporais recebem o **número de dias** contados a partir do diagnóstico. Veja Dicionário de dados (cap. 8).

O que é o "dia 0"?

É a data do diagnóstico, ou seja, do primeiro exame que concluiu pela malignidade. Todos os demais eventos são contados a partir dela. Por exemplo, uma amostra coletada 15 dias depois do diagnóstico deve ser informada com o valor `15` em `days_to_sample`.

Como converto as datas do prontuário em número de dias?

Use o **conversor de datas em dias**, disponível nos formulários de cadastro da plataforma. A conversão é feita no seu próprio computador, e as datas originais não devem ser enviadas ao ATPBR.

Existe alguma data que deve ser informada como data?

Sim, apenas a **data de sequenciamento** (`date_of_sequencing`), no formato DD/MM/AAAA, por não se referir ao paciente.

Identificadores e barcode

Quem define o `patient_id` e o `sample_id`?

O centro coletor. Ambos são códigos de 4 dígitos e não devem ter relação com dados pessoais: não use, por exemplo, o número do prontuário nem partes da data de nascimento.

Preciso preencher o barcode?

Não. O barcode é gerado automaticamente pela plataforma a partir dos campos informados no cadastro do paciente, da amostra e do experimento. Veja Sistema de identificação e barcode (cap. 9).

Por que aparece "XXX" no barcode?

O "XXX" indica que um dos campos usados na composição do barcode ainda não foi preenchido ou não pôde ser identificado. Confira os campos obrigatórios do registro.

Por que o número da corrida aparece como "00"?

O número da corrida é atribuído pela combinação de data de sequenciamento, centro de sequenciamento e plataforma de sequenciamento. Enquanto algum desses três campos não estiver preenchido, a corrida aparece como `00`.

Recebi um aviso de cadastro em duplicidade. O que significa?

A plataforma não aceita dois registros com o mesmo barcode. O aviso indica que já existe um paciente, uma amostra ou um experimento com os mesmos dados de identificação. Verifique se o registro não foi cadastrado anteriormente ou se o código informado está correto.

Amostras

Minha amostra foi recusada por falta de estadiamento. Por quê?

Cada amostra deve informar pelo menos um sistema de estadiamento (FIGO, TNM, Rai, Binet, LMC Phase, AML/ALL Stage ou CIN). Basta o sistema aplicável ao caso; os campos AJCC não atendem a essa exigência. Veja **Dicionário de dados** (cap. 8).

Onde informo o anticoagulante do tubo de coleta (EDTA, citrato)?

No campo `collection_tube_additive` (aditivo do tubo de coleta), que é opcional. Esse campo não faz parte da preservação da amostra e não entra no barcode.

Arquivos de sequenciamento

Em que formato devo enviar os arquivos de sequenciamento?

FASTQ compactado com gzip (`.fastq.gz` ou `.fq.gz`). Em leituras pareadas (paired-end), envie os dois arquivos de cada amostra. Veja **Tipos de dados** (cap. 10).

Por que não consigo cadastrar os experimentos antes de enviar os FASTQ?

Cada experimento indica, nos campos `fastq_1` e `fastq_2`, quais arquivos pertencem àquela amostra. Por isso, os arquivos precisam estar disponíveis no Ensaio antes do cadastro dos experimentos. Veja **Submissão de dados** (cap. 5).

Meus FASTQ já estão na Plataforma Multiusuário de Bioinformática do INCA. Preciso enviá-los de novo?

Não. Informe o caminho absoluto do diretório onde estão os arquivos no campo *Repositórios Associados* do Ensaio. Veja **Submissão de dados** (cap. 5).

Recebi um aviso de que um arquivo FASTQ não foi encontrado. O que significa?

O nome informado em `fastq_1` ou `fastq_2` não corresponde a nenhum arquivo do Ensaio. Confira se o arquivo foi enviado ao Ensaio correto e se o nome foi digitado exatamente igual, incluindo a extensão.

O nome do arquivo FASTQ pode conter o nome do paciente ou o número do prontuário?

Não. Os nomes de arquivo também não podem conter dados pessoais. Renomeie os arquivos antes do envio, se necessário.

Preciso enviar arquivos BAM, CRAM ou VCF?

Não. Esses arquivos são gerados pelos pipelines do ATPBR a partir dos FASTQ enviados. Veja **Plataforma analítica** (cap. 11).

Preciso executar algum pipeline?

Não. Os pipelines são executados centralmente pela equipe do ATPBR. Veja **Plataforma analítica** (cap. 11).

Governança e ética

O que é o Termo de Responsabilidade do Pesquisador Remetente?

É o termo que o pesquisador aceita antes de enviar dados ao ATPBR. Nele, declara que os dados são autênticos e foram desidentificados, que vai informar a documentação ética e as permissões de uso de cada conjunto e que vai comunicar mudanças, como retiradas de consentimento ou novas restrições. Veja **Submissão de dados** (cap. 5).

Meu estudo tem aprovação ética. Isso basta para compartilhar os dados no ATPBR?

Não necessariamente. A aprovação ética do estudo de origem não autoriza, por si só, novos usos ou compartilhamentos. É preciso verificar se o TCLE e o protocolo permitem a incorporação ao ATPBR e o uso futuro dos dados, e se a instituição de origem autoriza o compartilhamento. Veja **Governança de dados** (cap. 4).

O sistema aceitou meus dados. Eles já estão disponíveis para outros pesquisadores?

Não. Ser aceito pelo sistema não equivale a ser admitido. Cada conjunto passa por verificações técnicas e ético-documentais e só fica disponível depois de registrado como apto. Conjuntos pendentes ficam bloqueados até a regularização. Veja **Submissão de dados** (cap. 5).

Preenchi o campo de aprovação ética. Preciso enviar algum documento?

Sim. O campo preenchido não substitui o documento comprobatório. A equipe do ATPBR pode solicitar o parecer do CEP ou da Conep, a versão do TCLE (ou a decisão de dispensa de consentimento) e o instrumento de transferência de dados. Veja **Governança de dados** (cap. 4).

Um participante retirou o consentimento. O que devo fazer?

Comunique a equipe do ATPBR pelo e-mail atpbr@inca.gov.br, identificando o participante **apenas pelo código** usado na plataforma (`patient_id` ou barcode), nunca pelo nome ou documento. O ATPBR bloqueia novos usos dos dados e, quando necessário, avisa os pesquisadores que já os receberam. Veja **Governança de dados** (cap. 4).

O acesso aos dados inclui as amostras biológicas?

Não. O acesso a dados e o acesso ou a transferência de amostras são autorizações distintas. As decisões sobre amostras cabem às instituições que as guardam. Veja **Governança de dados** (cap. 4).

Acesso aos dados

Quais dados posso consultar sem solicitar autorização?

As informações aprovadas para divulgação pública: os metadados do catálogo de dados e os resultados publicados no **Navegador de resultados** (cap. 12) (cBioPortal do ATPBR).

Quem decide se os dados do meu projeto aparecem no cBioPortal?

O responsável pelo projeto indica, no cadastro de cada Ensaio, se os dados podem ser disponibilizados. A publicação, porém, só acontece depois que o conjunto é validado e que a divulgação é aprovada, com avaliação dos riscos de reidentificação e respeitando o consentimento dos participantes e as permissões de origem. Os FASTQ e CRAM permanecem sempre sob acesso controlado. Veja **Acesso aos dados** (cap. 6).

Como obtenho acesso a dados controlados, como FASTQ, CRAM ou variantes germinativas?

Registre a solicitação no módulo de pedidos de acesso da plataforma ATPBR, com login institucional (dúvidas: atpbr@inca.gov.br). A solicitação é analisada pelo Comitê de Acesso a Dados (CAD) e, se aprovada, a liberação depende da assinatura do instrumento de uso de dados. A autorização do CAD não substitui a aprovação ética da sua pesquisa. Veja **Acesso aos dados** (cap. 6).

Meus colaboradores podem usar o meu acesso?

Não. O acesso é individual e as contas não podem ser compartilhadas. Cada colaborador precisa ser identificado no pedido e aceitar as mesmas obrigações. Veja **Políticas de uso** (cap. 13).

Como devo citar o ATPBR?

Cite o ATPBR e os estudos de origem dos dados. O artigo de referência está em **Sobre o ATPBR** (cap. 1), e as regras de uso, em **Políticas de uso** (cap. 13).

Atlas Tumoral da População Brasileira

Ampliar a representatividade da população brasileira na pesquisa oncológica, disponibilizando recursos que contribuam para a descoberta de biomarcadores, o desenvolvimento de novas estratégias de diagnóstico e tratamento e a oferta de medicina de precisão no SUS.

SITE **www.atpbr.com.br**

CONTATO **atpbr@inca.gov.br**

Financiamento: Ministério da Saúde, em parceria com o CNPq

GENOMAS BRASIL INCA SUS MINISTÉRIO DA SAÚDE